Patient flow and congestion in the out-patient department at Zithulele Hospital

Emma Gibson

Thesis presented in fulfilment of the requirements for the degree of
Master of Science
in the Faculty of Science at Stellenbosch University

Supervisor: Professor SE Visagie

March 2017

Stellenbosch University https://scholar.sun.ac.za

Declaration

By submitting this thesis electronically, I declare that the entirety of the work contained therein is my own, original work, that I am the sole author thereof (save to the extent explicitly otherwise stated), that reproduction and publication thereof by Stellenbosch University will not infringe any third party rights and that I have not previously in its entirety or in part submitted it for obtaining any qualification.

Date: March 2017

Copyright c◯ 2017 Stellenbosch University

All rights reserved

Stellenbosch University https://scholar.sun.ac.za

Stellenbosch University https://scholar.sun.ac.za

Abstract

This thesis considers the causes and effects of congestion in the out-patient department (OPD) at Zithulele hospital, a rural healthcare facility located in the Eastern Cape province of South Africa. Detailed mathematical models are developed to analyse the flow of different types of patients through the OPD and evaluate strategies for reducing the negative effects of congestion. The OPD model is implemented as a decision support tool which can be used by hospital staff.

Stellenbosch University https://scholar.sun.ac.za

Stellenbosch University https://scholar.sun.ac.za

Acknowledgements

The author wishes to acknowledge the following people for their various contributions towards the completion of this work:

- Dr Ben Gaunt
- Dr Gareth Meyer

Stellenbosch University https://scholar.sun.ac.za

Stellenbosch University https://scholar.sun.ac.za

Table of Contents

Stellenbosch University https://scholar.sun.ac.za

Stellenbosch University https://scholar.sun.ac.za

Stellenbosch University https://scholar.sun.ac.za

Stellenbosch University https://scholar.sun.ac.za

List of Reserved Symbols

Symbols in this thesis conform to the following font conventions:		
$\mathcal{A}$	Symbol denoting a **set**	(Calligraphic capitals)
$\mathbf{A}$	Symbol denoting a **matrix**	(Roman bold capitals)
$\mathbf{a}$	Symbol denoting a **vector**	(Roman bold lowercase)
A, a, α	Symbol denoting a **variable or function**	(Italic capitals, lowercase and Greek letters)

Symbol	Meaning
$\times$	Symbol used to denote the multiplication operator
$\sum$	Symbol used to denote the summation operator
$\prod$	Symbol used to denote the product operator
min	Symbol used to denote the minimum operator
max	Symbol used to denote the maximum operator
$P(X \leq x)$	The probability that a random variable, X, takes on a value smaller than or equal to the value x.
$\int f(t)\ \mathrm{d}t$	The integral of a function, $f(t)$, with respect to t.
$\frac{\mathrm{d}f(t)}{\mathrm{d}t}$	The derivative of a function, $f(t)$, with respect to t.

Stellenbosch University https://scholar.sun.ac.za

Stellenbosch University https://scholar.sun.ac.za

List of Acronyms

HIV: human immunodeficiency virus

ARV: antiretroviral

TB: tuberculosis

NGO: non-governmental organisation

OPD: out-patient department

FCFS: first come–first serve

DES: discrete event simulation

OR: Operations Research

GA: genetic algorithm

Stellenbosch University https://scholar.sun.ac.za

Stellenbosch University https://scholar.sun.ac.za

List of Figures

Stellenbosch University https://scholar.sun.ac.za

Stellenbosch University https://scholar.sun.ac.za

List of Tables

Stellenbosch University https://scholar.sun.ac.za

Stellenbosch University https://scholar.sun.ac.za

CHAPTER 1

Introduction

The South African healthcare system consists of a large network of state-funded healthcare facilities, as well as a smaller private sector. Approximately 80% of the country's population is dependent on the state health care system, despite the fact that private health insurance accounts for 42% of South Africa's total healthcare expenditure (Day & Gray, 2016; World Health Organisation, 2016). Discrepancies between the private and public healthcare sectors reflect the alarmingly high levels of social and economic inequality in South African society.

Many of the challenges faced by the public healthcare sector are rooted in South Africa's long history of racial segregation, which systematically disenfranchised the country's black, coloured and Indian populations and de-prioritised the development of healthcare infrastructure for these groups. The new South African government has made some progress towards addressing these issues by consolidating the state healthcare system and investing heavily in the development of healthcare facilities in previously disadvantaged areas. However, many of these policies have fallen far short of their objectives due to poor implementation and leadership (Coovadia *et al.*, 2009).

The state's failure to address the spread of HIV during the period 1994–2004 was heavily criticised both locally and internationally, and is estimated to have resulted in the deaths of 330 000 people (Simelela *et al.*, 2015). Although there has been a dramatic improvement in the government's response to this issue over the last decade, access to antiretroviral (ARV) therapy and other life saving drugs is still unreliable due to supply-chain problems which result in frequent stock-outs (Bateman, 2013a). The nationwide epidemic of HIV/AIDS and associated illnesses such as tuberculosis (TB) places a heavy burden on the public health sector.

Although less than a fifth of South Africa's population has access to private healthcare, 55% of the country's doctors are employed in the private healthcare sector (World Health Organisation, 2016). Under-staffing in public healthcare facilities is a major challenge, and during the period 2002–2010 the percentage of vacant health professional posts in the public sector ranged from 27.2–42.5% (Health Systems Trust, 2016). Despite numerous government initiatives to address staff shortages, the poor working conditions in state healthcare facilities make it very difficult for these facilities to recruit and retain staff (Breier, 2007).

This is particularly true of rural hospitals, where factors such as geographical isolation, difficult living conditions, under-resourced facilities, and incompetent management lead to low staff morale (Bateman, 2013b). In addition to persistent staff shortages, rural facilities are also affected by overwhelming social challenges such as a lack of access to education, housing, sanitation, clean water, and adequate nutrition in the surrounding communities (Mayosi & Benatar,

Stellenbosch University https://scholar.sun.ac.za

2014). A report entitled *Death and dying in the Eastern Cape: An investigation into the collapse of a health system* (Thom *et al.*, 2013) highlights the shocking conditions that exist in many rural hospitals and the daily struggles of healthcare workers and patients in these facilities. In spite of these difficult conditions, there are a number of dedicated healthcare workers who are firmly committed to improving healthcare in rural areas.

1.1 Zithulele Hospital

Zithulele Hospital is located amongst a collection of rural villages and homesteads in the Transkei region of the Eastern Cape. The hospital was founded in 1956 by missionaries of the Dutch Reformed Church, and became part of the state healthcare system in 1976. It currently falls under the jurisdiction of the Eastern Cape Department of Health.

Zithulele faces many of the operational challenges that are common in rural hospitals: under-funding, incompetent administration, poor infrastructure, drug shortages and a lack of access to important medical resources and equipment (Baleta, 2009; Gaunt, 2010; Young & Gaunt, 2014). The hospital has also struggled to retain staff, and during certain periods of Zithulele's history there was not a single doctor stationed at the hospital.

Historically, the high staff turnover rate at Zithulele has made it very difficult to address the long term development of the hospital. Short term staff focussed most of their attention on the immediate challenges that they encountered during their brief period at Zithulele, and the cycle of constantly changing staff resulted in a lack of consistent management.

A major turning point was the arrival of Drs Ben and Taryn Gaunt, who began working at Zithulele in July 2005. Recognising the need for leadership at the hospital, Ben and Taryn moved their family to Zithulele and made a long term commitment to improving the standard of healthcare in the area.

Ben and Taryn's decision had a profound impact on morale at Zithulele and helped to provide a sense of direction and hope for the hospital's future. Inspired by their example, many other medical professionals have joined the hospital's staff and Zithulele now has the equivalent of fourteen full-time doctors. Zithulele is regularly inundated with requests for medical internships and provides opportunities to many young doctors to learn about the challenges involved in rural healthcare.

Between 2005 and 2008, a number of steps were taken to improve the hospital's maternity service. Deliveries in the hospital rose by 53% during this period, and the perinatal mortality rate dropped from 49.1 per 1000 in 2005 to 22.4 per 1000 in 2008 (Gaunt, 2010). There was also a significant improvement in strategies for the prevention of mother-to-child transmission (PMTCT) of HIV through increased emphasis on HIV testing. Nearly 60% of the women who gave birth at Zithulele in 2006 did not know their HIV status, but buy 2008 this number was reduced to 0.2% (Gaunt, 2010).

Another of the hospital's major achievements over the past decade was the implementation of highly active antiretroviral therapy (HAART), a new HIV treatment model. Between 2005 and 2013, over 5000 patients were initiated on antiretrovirals through a network of local hospitals and clinics, and no patient in this program was ever turned away without receiving their full treatment regimen (Young & Gaunt, 2014).

Stellenbosch University https://scholar.sun.ac.za

1.1.1 Overview of hospital services[1]

Zithulele is a district hospital that provides healthcare services to a population of about 130 000 people from villages and homesteads in the surrounding areas. The hospital is also involved in an extensive network of community outreach programs in conjunction with local NGOs.

The hospital staff assist at local clinics, which provide basic care and educational services at a community level. These clinics help to reduce the patient load at the hospital and are easily accessible to patients who cannot afford to travel to the hospital to seek treatment. The clinics address a wide range of medical needs such as basic dental care, optometry, physical therapy, rehabilitation, diabetic clinics, and nutritional advice.

Zithulele's in-patient facilities include general wards as well as specialist wards for paediatric cases, TB patients, and maternity patients. Caesarean sections and minor surgeries are performed at the hospital, while more complicated surgical cases are transported to larger hospitals in Mthatha or East London. The hospital also offers a number of onsite diagnostic tests including basic haematology, X-rays and ultrasounds.

The bulk of the hospital's patients are treated in the out-patient department (OPD). The OPD provides access to generalist doctors, diagnostic tests, and a dispensary which supplies drugs prescribed for both chronic and acute conditions. The OPD facilities also include a number of other health-related services such as physiotherapy, dentistry, optometry, audiology, occupational therapy, nutritional counselling, psychology, psychiatry, and trauma counselling. Due to resource constraints, Zithulele does not have separate casualty facilities and all casualty patients are treated in the OPD.

1.1.2 Problem identification

There has been a significant increase in the number of patients treated at Zithulele over the last decade, particularly in the OPD. Figure 1.1 shows the annual OPD patient counts, which have more than tripled since 2005. The scope of treatments available at the hospital has also increased dramatically due to the arrival of specialist staff and their efforts to secure access to new equipment and resources for the hospital.

Despite these advances, the hospital still faces significant challenges. Important upgrades to the hospital's infrastructure have been delayed by several years due to a lack of funding, and current budget constraints often prevent the hospital from procuring medical equipment and supplies. Access to the hospital is still a major issue, since many patients struggle to afford the cost of transport to and from the hospital in local minibus taxis. This is particularly problematic for patients on long term treatments such as ARVs, and some staff at Zithulele help to cover these costs out of their own pockets.

One of the current concerns that the hospital faces is the high level of congestion in the OPD. Queues of patients begin forming outside the OPD early in the morning and often continue throughout the day, resulting in long delays for patients seeking treatment. In some cases, patients who live far away are forced to wait at the hospital overnight because they cannot afford to travel home and return to the hospital the next day.

Doctors at Zithulele are particularly concerned about the effect of this congestion on high-risk groups such as casualty patients, elderly patients, maternity patients and children. Since these patients are often caught up in long queues, serious injuries or illnesses may not be identified

[1] Based on information available from the Zithulele Hospital website (2016) .

Stellenbosch University https://scholar.sun.ac.za

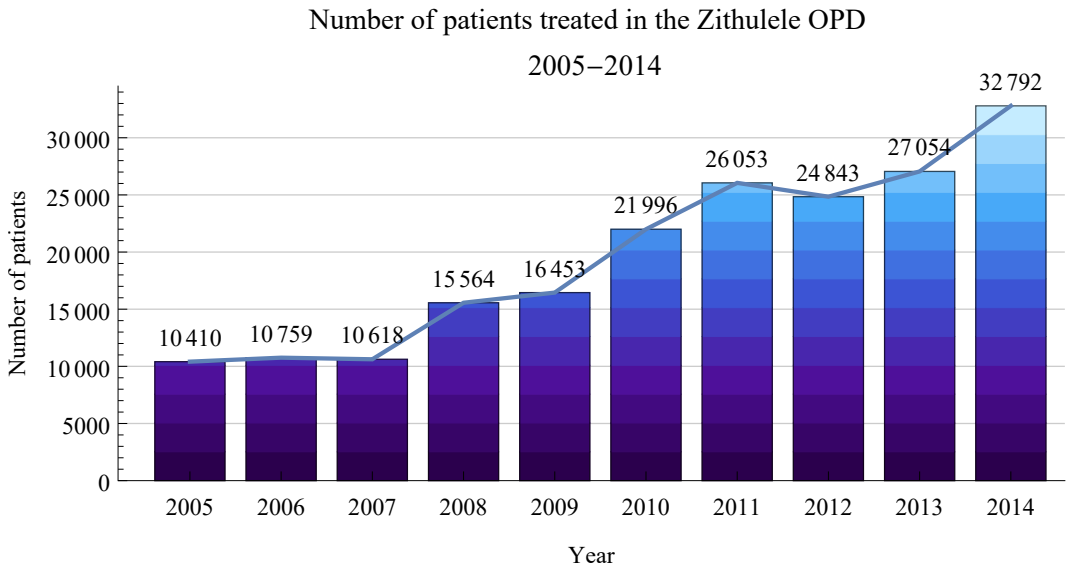


FIGURE 1.1: *A plot of the annual number of patients treated in the Zithulele OPD from 2005 to 2014, based on statistics reported on the Zithulele Hospital website (2016).*

until the patient's condition deteriorates enough to attract the attention of the OPD staff. The long queues are also problematic for patients with weaker immune systems, who are more likely to contract infections such as TB or influenza from other patients in the queue.

The congestion in the OPD leads to a high-pressure, chaotic work environment. Doctors and other hospital staff are forced to move through patients as quickly as possible, which increases the chances that they will make errors or miss important signs and symptoms. This is particularly problematic for doctors who work with translators, and can result in a patient leaving the hospital without a proper understanding of their treatment.

Efforts to reduce the congestion in the OPD are hampered by the complexity of the OPD system. During visits to the OPD, patients are usually required to stand in multiple different queues associated with various administrative and medical processes. Very little is understood about the interactions between the queues at these different processes and their effect on patient waiting times.

The diverse mix of patients in the OPD adds an additional level of complexity to this problem. Different types of patients have a wide range of treatment needs, and balancing these needs is an important part of improving the overall standard of care in the OPD. Although different types of patients are mixed together in many of the OPD queues, they are not all affected equally by the congestion in the OPD system.

The initial concern that will be addressed in this thesis is the need for a more comprehensive understanding of the congestion in the OPD. The scope of this problem includes a detailed analysis of the different OPD queues, as well as the different types of patients in system. Rather than considering these components of the system in isolation, this research focuses on the interactions that take place between different components of the OPD system.

This thesis also considers how the negative effects of congestion in the OPD can be reduced. Both short term and long term strategies are needed to improve the efficiency of the facility, especially if the number of patients continues to increase over the next few years. The scope of this problem considers the hospital's financial constraints, as well as the limitations associated with the OPD's location, staff, and infrastructure.

Stellenbosch University https://scholar.sun.ac.za

1.2 Aims and objectives

Aim 1: Understanding patient flow in the OPD

Objective 1.1: Develop detailed mathematical models to describe patient flow in the OPD.

Objective 1.2: Analyse the causes and effects of congestion in the OPD.

This thesis aims to develop a better understanding the OPD queueing process and its influence on the flow of different types of patients through the system. Although there is a general consensus that the queues in the OPD are too long, it is quite difficult to identify all the factors that contribute to congestion in the system. More insight is needed to understand why congestion occurs and how it affects patients in the OPD.

Aim 2: Strategies for improving the OPD

Objective 2.1: Evaluate strategies to address the causes of congestion.

Objective 2.2: Evaluate strategies to mitigate the negative effects of congestion.

This thesis considers various strategies to address both the causes and effects of congestion in the OPD. The aim of this analysis is to identify inexpensive ways to improve the efficiency of the facility and ensure that patients' needs are met.

Aim 3: Practical implementation

Objective 3.1: Develop a decision support tool to give hospital staff access to models and results.

There is a strong emphasis on the practical relevance of this thesis and its contribution to the ongoing efforts raise the standard of care in the OPD. The research presented in this thesis aims to provide insights and suggestions that will assist staff in improving the efficiency of the OPD's facilities.

1.3 Thesis layout

The first part of this thesis focusses on modelling the OPD queueing system. Chapter 2 introduces a conceptual model which captures the structure of the OPD system and provides a detailed description of its different components. The purpose of this conceptual model is to define the notation and parameters that are used in the mathematical models in Chapters 3 and 4.

Three queueing theory models for the OPD system are developed in Chapter 3. These queueing theory models focus on the length of the OPD queues at different times during the day and their impact on the overall flow of patients through the system. Chapter 4 introduces an agent-based simulation model, which focuses on individual patients in the OPD and the effects of congestion on their progress through the OPD.

Stellenbosch University https://scholar.sun.ac.za

Chapter 5 discusses the input data that is needed to model different components of the OPD system and describes how various model parameters were estimated from data collected in the OPD. This chapter also considers the challenges posed by the lack of reliable data concerning certain elements of the OPD system and discusses the some of the difficulties associated with data collection.

The results of the simulation and queueing theory models are compared in Chapter 6, and a detailed discussion of the advantages and disadvantages of each model is provided. A sensitivity analysis is conducted to determine how certain model parameters affect the simulation results.

Chapter 6 also uses the simulation model to illustrate some of the changes that were implemented in the OPD during 2015. The consequences of these changes are discussed in terms of their effect on the length of the OPD queues, as well as patient waiting times.

Chapter 7 introduces a genetic algorithm for improving the distribution of staff members in the OPD. The genetic algorithm generates alternative staff schedules by considering the number of and type of staff available. The optimisation algorithm is tested using the 2015 OPD staff schedules, and three alternative schedules are compared to the original OPD schedule.

Chapter 8 focuses on the practical implementation of the OPD model at Zithulele. This chapter describes the development of a decision support tool that allows the OPD staff to investigate the causes of congestion in the OPD system and its impact on different types of patients. The decision support tool can also be used to test various strategies for reducing the effects of congestion.

The thesis concludes in Chapter 9 with a discussion of some of the insights gained into the queueing process in the OPD and its effects on different types of patients. Chapter 9 also considers how these insights can be used to improve the efficiency of the OPD, as well as opportunities for further work.

Stellenbosch University https://scholar.sun.ac.za

CHAPTER 2

Conceptual model and notation

This chapter introduces a conceptual model for the queueing system in the Zithulele OPD, which will provide notation for the mathematical models in Chapters 3, 4 and 7. The OPD conceptual model consists of two key components: processes and patients. The important characteristics of these components and their relevance to the model are discussed in § 2.1–§ 2.2, and § 2.3 describes the interactions between these components in the OPD model. Some of the explicit and implicit assumptions of the model are discussed in § 2.4.

2.1 Modelling OPD processes

In the OPD conceptual model, a *process* is any treatment or service that is provided to patients in the OPD. This includes a variety of administrative, diagnostic and medical procedures. In the mathematical models, the OPD processes are represented by the set $\mathcal{I} = \{1, 2, \ldots, n\}$. Individual processes are also given descriptive names (written in UPPERCASE) based on their role in the system.

Since it would be impractical to create an exhaustive list of every medical procedure performed in the OPD, this definition can be simplified by grouping all procedures that are associated with the same queue into a single process. For example, all patients wait in the same queue to see a doctor, regardless of the nature of their injury or illness, so consultations with the OPD DOCTORS are considered to be a single process.

Other examples of processes in the OPD include:

- CLERKS: Staff record the patient's details and stamp their patient book.
- VITALS: Nurses check the patient's blood pressure, heart-rate, breathing and weight.
- TRIAGE: A preliminary assessment to determine the type of injury/illness and the severity of a patient's condition.
- X-RAYS, BLOOD TESTS: Tests required to diagnose and/or monitor various injuries and illnesses.
- PHARMACY: Patients collect medication prescribed for chronic and acute conditions.

The processes operate as a network, with patients moving from one process to another throughout the day. Each process has a different role to fulfil and may depend on one or more of the

Stellenbosch University https://scholar.sun.ac.za

other processes in the system. If the system is working efficiently, patients should be able to access each process without experiencing long delays and there should be a steady flow of patients from one process to another.

2.1.1 Staff

Often, multiple staff members are assigned to the same process so that several patients can be treated concurrently. Each process has its own staff schedule which is given by a step function, $\varsigma_i(t)$, that indicates the number of staff on duty at process $i \in \mathcal{I}$ at time t.

The OPD conceptual model assumes that each staff member can help exactly one patient at a time, and that all staff members operate independently of each other. In cases where this assumption does not hold, $\varsigma_i(t)$ is adjusted to reflect the number of patients that can be accommodated simultaneously, rather than the number of staff.

2.2 Modelling OPD patients

Given that the OPD provides so many different medical services, there are many different types of patients that come to the facility. It is important to differentiate between patient types in the model, since certain patients place a higher demand on the network and are more sensitive to long waiting times. Rather than trying to model an "average patient", the OPD conceptual model incorporates a number of distinct *patient profiles* to reflect theses different types of patients. The patient profiles are given by the set $\mathcal{P} = \{1, 2, \ldots, m\}$.

Patients in the OPD usually fit into one of two broad classes:

- Returning patients:
  Returning patients are treated for long term or chronic conditions such as TB or HIV. These patients tend to come to the OPD on a monthly basis for monitoring and medication.
- Casualty patients:
  Casualty patients can have a wide variety of injuries and illnesses, with varying levels of severity. These patients generally require once-off treatment for acute conditions.

These two patient groups can be divided into a number of more detailed sub-groups by considering the types of treatments that are administered to different patients.

2.2.1 Waiting time targets

Although the OPD aims to treat all patients as quickly as possible, the consequences of delaying treatment are a great deal worse when a patient is critically ill or injured. The patient profiles are associated with a set of waiting time targets,

$$\mathcal{W} = \{w_1^{(u)}, w_2^{(u)}, \ldots, w_m^{(u)}\}, \tag{2.1}$$

which indicates the maximum amount of time that a patient in each profile should be expected to wait for treatment.

These targets are determined by the hospital's own standards, as well as guidelines from the Department of Health. The percentage of patients who are treated within these targets is considered to be an important measure of the quality of service in public healthcare facilities.

Stellenbosch University https://scholar.sun.ac.za

2.2.2 Number of patients

The total number of patients treated in the OPD in a day is given by the summation

$$\eta = \sum_{p \in \mathcal{P}} \eta_p, \tag{2.2}$$

where η_p is the number of patients from each different profile. Patient arrivals can vary greatly from day to day, and so the daily patient count for each profile is modelled as a random variable with the probability mass function

$$f_{\eta_p}(x) = \mathrm{P}\left(\eta_p = x\right), \qquad \text{with } x \in \{\eta_p^{(l)}, \eta_p^{(l)} + 1, \ldots, \eta_p^{(u)}\}. \tag{2.3}$$

The lower and upper bounds $\eta_p^{(l)}$ and $\eta_p^{(u)}$ represent the minimum and maximum number of patients from each profile that could arrive at the OPD on any particular day, and $f_{\eta_p}(x)$ is estimated from the data discussed in Chapter 5.

2.2.3 Arrival times

The OPD does not schedule appointments, so patients decide when to come to Zithulele based on the availability of transport and the urgency of their condition. Different types of patients tend to follow different arrival patterns. For example, many returning patients plan their trips in advance and get to the hospital early, while casualty patients are likely to arrive randomly throughout the day.

The distribution of new arrivals over the course of a day is modelled by a set of probability density functions $\alpha_p(t)$, with $p \in \mathcal{P}$. There are two ways to interpret these arrival functions:

- Consider an individual patient from profile p who arrives at the OPD on a certain day. The probability that their arrival will occur in a specific interval $[t_1, t_2]$ is given by the integral $\int_{t_1}^{t_2} \alpha_p(t)\mathrm{d}t$.
- Consider all the patients from profile p who arrive at the OPD on a certain day. The function $\eta_p \alpha_p(t)$ is the average rate at which new arrivals are occurring at time t.

2.3 Modelling OPD interactions

This section introduces a set of treatment parameters that describe how patients in each profile interact with the OPD processes. These interactions depend on a number of factors, including the OPD administrative procedures, the specific needs of each patient, and the urgency of their condition.

2.3.1 Treatment needs

The treatment needs of each patient profile determine which processes they undergo while they are in the OPD. Treatment needs are represented by a set of $n \times m$ parameters

$$\varrho_i^p, \qquad \text{with } i \in \mathcal{I} \text{ and } p \in \mathcal{P}, \tag{2.4}$$

which give the probability that a patient from profile p is will need process i.

Stellenbosch University https://scholar.sun.ac.za

In some profiles, not all patients will follow an identical route through the OPD. Differences in individual cases may allow some patients to skip certain processes, or make it necessary to complete certain additional processes. The parameters ϱ_i^p may therefore take on any value between 0 and 1.

2.3.2 Routing

Patients move through the OPD processes in a specific order, since most process cannot be completed without gathering all the necessary information from previous steps. This order is represented by a set of routing parameters

$$\phi_i^p \in \{1, 2, \ldots, n\}, \qquad \text{with } i \in \mathcal{I} \text{ and } p \in \mathcal{P}. \tag{2.5}$$

The routing parameters rank the different OPD processes in the order in which they are visited, so $\phi_i^p = 1$ indicates that patients from profile p begin at process i, while $\phi_i^p = n$ indicates that patients from profile p leave the OPD after process i.

It is necessary to create separate rankings for each profile because this order is not the same for different types of patients. For example, patients returning for a check-up can have BLOOD TESTS or X-RAYS before they go to the DOCTORS, while patients who arrive with a new injury or illness will need to go to the DOCTORS first and then proceed to the appropriate tests.

2.3.3 Treatment times

In the OPD conceptual model, *treatment times* refer to the amount of time needed to treat a patient, rather than the time at which the treatment takes place. The treatment time for patients from profile p at process i is given by the parameter τ_i^p.

Exact treatment times vary for individuals within each profile, especially for urgent cases. The treatment times are therefore modelled as random variables with the probability density function

$$f_{\tau_i^p}(x), \qquad \tau_i^{p(l)} \leq x \leq \tau_i^{p(u)}, \tag{2.6}$$

where $\tau_i^{p(l)}$ and $\tau_i^{p(u)}$ represent the lower and upper bounds for treatment times and $f_{\tau_i^p}(x)$ is estimated from the data discussed in Chapter 5.

2.3.4 Priority

The priority parameters describe the order in which different patient profiles are treated at each process. Priorities are given by the variables

$$\vartheta_i^p \in \{1, 2, \ldots, m\}, \qquad \text{with } i \in \mathcal{I} \text{ and } p \in \mathcal{P}, \tag{2.7}$$

where $\vartheta_i^p = 1$ indicates that profile p patients have first priority at process i. It is possible for two or more profiles to have the same priority (i.e. $\vartheta_i^{p_1} = \vartheta_i^{p_2}$), which means that patients from these profiles are treated on a first-come, first-serve basis.

Profiles with a higher priorities are treated before other patients in the queue, regardless of who arrived first. If there are multiple different priority levels in one queue, then staff will begin with the highest and work their way down to the lowest.

Urgent patients and patients who have waited overnight are often assigned a higher priority to allow them to see a doctor as quickly as possible. Unfortunately, this does not guarantee immediate treatment, as there are not always enough staff available to deal with these cases.

Stellenbosch University https://scholar.sun.ac.za

2.4 Assumptions

One of the biggest challenges in modelling the OPD system is the "human element". Individual OPD staff members and patients may sometimes behave in unexpected ways, which means that every treatment interaction that takes place in the OPD is slightly different. One way to deal with these discrepancies is to incorporate some level of variability in the OPD model by using random variables and probabilities. However, too much variability in the model makes it difficult to obtain a clear picture of the causes and effects of congestion within the system. To eliminate unnecessary noise and make the model's results more interpretable, certain assumptions are made about the behaviour of staff and patients in the OPD.

2.4.1 Treatment assumptions

The parameters in § 2.3 allow for some variability in the treatment interactions between staff and patients, such as different treatment times. The source of this variability is the patient profiles, and the differences between individual patients within each profile. The treatment parameters do not include any variability associated with the OPD staff, which leads to two important assumptions:

Assumption 1: Staff work at the same speed.

The OPD conceptual model assumes that all the staff at a given process are equally efficient. This means that a patient's treatment time will not be influenced by which staff member they see, and depends entirely on their medical needs.

Assumption 2: Treatment times are directly related to patients-per-hour.

It is assumed that staff do not take unscheduled breaks, and that the treatment times include any time needed to change between different patients. This means that if the average treatment time is 10 minutes per patient, staff are expected to see about 6 patients per hour.

2.4.2 Congestion assumptions

An underlying assumption of this model is that the congestion in the OPD is entirely dependent on the behaviour of staff and patients. There is also an implicit assumption that this is a one-way relationship — that the state of the queues does not influence the behaviour of patients and staff — since none of the model parameters allow for different behaviours when the OPD is very busy. This leads to the following assumptions:

Assumption 3: Patient arrivals are not affected by congestion.

This assumption implies that the state of the OPD queues has no effect on the number of patients that arrive at the OPD, or their arrival times. It is based on the fact that Zithulele's patients are spread over a wide geographical region and have limited access to transportation. Most of the patients who come to Zithulele rely on taxis that follow regular schedules, so they tend to have little control over their arrival time.

Assumption 4: Patients who arrive at the OPD will remain there until they are treated.

This assumption implies that patients will not leave the OPD without completing all of the necessary processes, regardless of how long they need to wait. It is a fairly realistic assumption, since there are no other hospitals near Zithulele where patients could seek

Stellenbosch University https://scholar.sun.ac.za

treatment. Few patients can afford additional trips to the hospital — to the extent that patients often wait overnight for test results, instead of returning the next day — so patients are unlikely to leave the OPD without seeking treatment.

Assumption 5: Treatment times are independent of queue length.

This assumption is less realistic than the two previous assumptions, because it is likely that staff members work more quickly when there are long queues and slow down when there are fewer patients. Since there is generally a high level of congestion in the OPD, the OPD model assumes that staff only spend as long as they need to with each patient.

2.5 Summary

The conceptual model presented in this chapter describes the OPD system in terms of two important components, *processes* and *patient profiles*. The OPD processes are given by the set $\mathcal{I} = \{1, 2, \ldots, n\}$, and each process is associated with a staff schedule $\varsigma_i(t)$. The patient profiles $\mathcal{P} = \{1, 2, \ldots, m\}$ are used to represent the different types of patients that are treated in the OPD. The daily patient count and arrival times for each profile are given by the probability distributions $f_{\eta_p}(x)$ and $\alpha_p(t)$.

The structure of the OPD system is modelled by a set of treatment parameters, which describe the interactions between the different patients and processes in the OPD. The treatment needs of different patients are given by the parameters ϱ_i^p, which indicate the proportion of profile p patients who need to be seen at process i. The parameters $\varnothing_i^p$ indicate the order in which different patients visit each process, and the treatment times are modelled as random variables with probability distributions $f_{\tau_i^p}(x)$. The parameters ϑ_i^p assign a priority ranking to different patient profiles at each process, which allows high priority patients to move to the front of certain queues.

Stellenbosch University https://scholar.sun.ac.za

CHAPTER 3

Queueing theory models

In this chapter, the OPD system is approached from a queueing theory perspective. A review of some relevant literature is provided in § 3.1, which highlights certain important properties of the OPD system. Some general notation is introduced in § 3.2, and three mathematical models for the OPD queues are presented in § 3.3–§ 3.5.

3.1 Literature

Queueing theory has been studied extensively for many decades and a vast literature is available on the subject. This section contains only a brief outline of important queueing theory concepts that are relevant to the OPD system. Detailed, systematic introductions to many of these ideas can be found in textbooks such as Kleinrock (1975), White (2012), and Chan (2014).

The aims of this section are to **(a)** introduce some basic notation and terminology to frame the OPD model in terms of general queueing theory conventions; and **(b)** draw attention to existing models and results that are relevant to the problem under consideration.

3.1.1 Classification of queueing systems

The term *queueing system* refers to a single queue of customers/jobs that all need to access the same service. Different types of queueing systems are classified using Kendall's notation, which describes the arrivals, service times and the number of servers at a particular queue using a three character string: A/B/c.

The simplest example of this notation is an M/M/1 queue, which has exponentially distributed inter-arrival and service times and a single server. The letter M refers to the Markov (memoryless) property of the exponential distribution, which forms the basis of traditional queueing theory analysis (Chan, 2014). When queueing systems do not have exponentially distributed inter-arrival and service times, a number of other statistical distributions may be used to model these events. For example, the letter G is used to represent a general distribution, while D indicates deterministic arrivals or service times.

In queueing systems with multiple servers ($c > 1$), it is generally assumed that the servers operate independently and in parallel. The A/B/c notation also implies that there is no limit on the number of customers or jobs that may be in the queue at the same time. If the maximum queue length is restricted, the notation A/B/c/x is used.

Stellenbosch University https://scholar.sun.ac.za

Another important characteristic of a queueing system is the *queueing discipline*, which determines the order in which customers/jobs will be serviced. In many queueing systems, jobs are processed in the order in which they arrived and this is referred to as a first-come, first-served (FCFS) or first-in, first-out (FIFO) policy. Other service disciplines include last-come, first-served, random ordering, and priority ordering.

The first model in this chapter (§ 3.3) represents the OPD processes as M/M/c/x queues with priority ordering. The second model (§ 3.4) also uses the priority queueing discipline, but approximates the stochastic, multi-server queues with deterministic D/D/1 queues. The third model (§ 3.5) considers the same D/D/1 queues with FCFS service.

3.1.2 Non-stationary queues

Much of the existing queueing theory literature is dedicated to the analysis of queueing systems in which the arrival distributions, service distributions and the number of servers remain constant over time Green *et al.* (2007). In such cases, it is generally sufficient to analyse the average or long term behaviour of the queue.

A queue is called *non-stationary* if the arrival and/or service distribution or the number of servers changes over time. For non-stationary queues, a subscript is added to Kendall's notation to indicate which elements are time dependent, for example $A_t/B_t/c_t$.

There are a number of techniques that are used to modify non-stationary queues so that traditional (stationary) analysis methods can be applied. The simplest approach to these systems is to average time dependent parameters over the analysis interval, which is known as a simple stationary approximation (SSA). This approach is very easy to implement, but performs poorly in systems with arrival rates that fluctuate by more than 10% (Green *et al.*, 1991).

More detailed approximation methods include pointwise stationary approximations (PSA) and the average stationary approximation (ASA) proposed by Whitt (1991). These approaches divide the analysis period into a series of stationary queues with parameters based on point estimates (PSA) or averages over intervals proportional to the mean service time (ASA). Based on empirical results in Green & Kolesar (1991) and analytical results in Whitt (1991), these models can incorporate greater variations in the arrival rate than SSA models, and their accuracy improves with larger arrival and service rates.

These techniques form the basis for many staff scheduling algorithms such as the stationary independent period by period (SIPP) approach. These algorithms use performance targets such as maximum queue length/waiting times to calculate how many staff (servers) should be assigned to a particular queue with non-stationary arrival rates. Unlike the general ASA and PSA methods, scheduling algorithms generally divide the analysis period into regular intervals that align with the shifts and breaks in the staff schedules.

Despite the popularity of the SIPP approach, many authors have shown that it is unreliable because it fails to account for the dependence of the queue length and waiting times between different intervals. Improvements and alternatives to this algorithm include the lagged SIPP approach in Green *et al.* (2003), and the effective arrival rate approximation (EAR) in Thompson (1993).

Stolletz (2008) notes that all of these approaches involve restrictions which make them unsuitable for modelling busy systems. In queueing theory, the "busy-ness" of a system is described by the *traffic intensity*, which is usually denoted by the symbol ρ. Traffic intensity is defined as the ratio of the rate at which new work is arriving at a queue relative to the capacity of the servers

Stellenbosch University https://scholar.sun.ac.za

to process the work.

Most queueing theory analysis is restricted to systems with a traffic intensity lower than 1, since a stationary system with $\rho > 1$ develops an infinitely long queue. These queues do not have an equilibrium state, so performance measures such as the expected queue length or average waiting time cannot be calculated. In non-stationary systems, however, periods of high traffic intensity ($\rho(t) > 1$) do not necessarily lead to infinitely long queues. These periods can be offset by periods of lower traffic intensity, where staff process the backlog of work that builds up during peak times. In these types of queues, stationary approximation methods can only be applied to periods where the traffic intensity is lower than 1.

Given that some of the Zithulele OPD processes experience periods of very high traffic intensity, stationary approximation models are not appropriate for these queues. However, the models in § 3.3 and § 3.5 do apply stationary approximation techniques to simplify certain calculations. In these models, the non-stationary arrival rates are replaced with ASA piecewise-constant functions that match the intervals on the staff schedule.

There are many other methods for analysing non-stationary queues that are more flexible than stationary approximations, particularly for periods of high traffic intensity. The models in this chapter are based on two of these methods: the non-stationary Chapman-Kolmogorov equations, and deterministic fluid approximations.

The non-stationary Chapman-Kolmogorov equations are considered an "exact" method for non-stationary queues, and often used as a benchmark for evaluating the results of heuristic methods (Green & Kolesar, 1991; Ingolfsson *et al.*, 2007; Odoni & Roth, 1983). The Chapman-Kolmogorov equations are usually solved numerically, since analytical solutions are difficult to find, even for simple models with stationary parameters (Tipper & Sundareshan, 1990).

Reibman & Trivedi (1988) investigates three numerical methods for solving these systems, and concludes that the uniformisation approach — approximating the continuous-time system with a discrete-time Markov process — is generally more accurate and efficient than explicit numerical schemes like the Runge-Kutta method. However, uniformisation methods are not recommended for stiff systems, which require stable implicit numerical schemes.

Unfortunately, the Chapman-Kolmogorov equations for the OPD queues result in a system that is too large to be solved within reasonable time constraints — a realistic representation of the OPD system with 6 processes and 6 patient profiles requires approximately 2.6×10^{35} equations. Although this model cannot been implemented, the equations discussed in § 3.3 provide some insight into the structure of the OPD system.

The second and third models in this chapter (§ 3.4–§ 3.5) are based on fluid approximation methods, which were first proposed by Kurtz (1970, 1971). These models approximate discrete events and individual OPD patients with a continuous flow of patients between the OPD queues. This flow is governed by deterministic equations which are based on the mean inter-arrival and service times.

By eliminating the stochastic components of these parameters, fluid approximations are able to model the expected length of a queue without first finding the probability distribution for the queue length. Markov models require a separate solution for every possible state (in this case, every possible queue length), but fluid models condense this system of equations into a single, first-order differential equation.

The main advantage of fluid approximation models is their simplicity, which makes them easy to implement and computationally inexpensive. However, these simpler results have a much smaller scope for interpretation. Since fluid models only predict the mean behaviour of the queue, they

Stellenbosch University https://scholar.sun.ac.za

cannot be used to calculate performance measures related to the variance or uncertainty in queue length or waiting time.

The simplicity of the fluid model also has certain disadvantages in terms of accuracy. Fluid models are very sensitive to traffic intensity and are not appropriate for systems where the traffic intensity is consistently lower than 1 (Kivestu, 1976). The fluid models in § 3.4 and § 3.5 incorporate periods of both high and low traffic intensity, but the main focus of these models is the behavior of the OPD queues during periods of high traffic intensity. When the traffic intensity is low, the fluid models are still able to describe the flow of patients through the facility, but they do not give an accurate representation of the length of the individual queues.

3.1.3 Queue networks

Queueing networks are groups of multiple linked queueing systems where each queue is processed by an independent set of servers, but customers or jobs may move between different queues. Queue networks which allow customers to enter and leave the network are known as open networks, while closed networks only allow customers to move from one queue to another within the network.

An important concept in the analysis of queueing networks is the *routing matrix*, which describes how customers move through the queues in the network. A network of n queues requires an $n \times n$ routing matrix, $\mathbf{R}$, where each element $r_{j,i}$ gives the conditional probability that a customer exiting queue j will move to queue i. If the network routing is deterministic, then each element $r_{j,i}$ is either 0 (if customers never go from j to i) or 1 (if customers always go from j to i). In closed networks, the row sums of $\mathbf{R}$ must equal 1 to ensure that all customers stay in the network, i.e. $\sum_{i \in \mathcal{I}} r_{j,i} = 1$. Open networks (such as the OPD queues) may have $\sum_{i \in \mathcal{I}} r_{j,i} \leq 1$.

Networks of queues are often called *Jackson networks* in reference to Jackson (1957), which describes an open network of multi-server queues with exponentially distributed service times and Poisson arrivals. In Jackson (1963), this model is extended to cases where the arrival and service rates depend on the length of each queue. Jackson shows that each queue in these networks can be modelled as an independent system with Poisson arrivals and departures, and that the steady-state distribution of the system as a whole can be expressed as a product of the Poisson flows at each node.

A broader class of queueing networks (known as *generalised Jackson networks*) includes systems with other arrival and service distributions. It is generally not possible to derive analytical results to characterise the steady-state distribution of queues in these networks, but several approximation methods can be used to analyse traffic through such systems. A recent review of this work can be found in Chen & Yao (2013).

The fluid approximation approach is a useful way to analyse the mean flow of traffic through queueing networks. This technique was initially used to model single-queue systems, and was later extended to stationary networks of queues in Newell (1982). The fluid models in § 3.4–§ 3.5 are based on the simple fluid approximations for networks with non-stationary arrival distributions in Tipper & Sundareshan (1990), and models for networks with non-stationary arrival distributions and servers in Liu & Whitt (2011).

Stellenbosch University https://scholar.sun.ac.za

3.1.4 Multi-class networks

Queues or networks that service multiple types of customers are known as multi-class systems. In these systems, different types of customers may be distinguished by a variety of properties, including arrival distributions, service-time distributions, queue disciplines, and routing policies. Multi-class systems are often indicated with the notation $\sum$A/B/c.

Multi-class networks were first introduced in Kelly (1975, 1976) as an extension of the Jackson network model. Kelly networks maintain the useful analytical properties of Jackson networks when arrivals and service times follow a Poisson distribution. Early work on approximation methods for multi-class queues includes aggregation techniques (Albin, 1982; Whitt, 1983b) which allow more general multi-class networks to be approximated as a single-class network. However, many authors have concluded that this approach is unreliable because it fails to account for the way that different classes interact within the same queue (Bitran & Tirupati, 1988; Whitt, 1983a).

Fluid approximation models for networks with non-stationary arrival rates are proposed in Tipper & Sundareshan (1990) and Sharma & Tipper (1993). In these models, different customer classes are represented by separate streams of fluid, which are channelled through a shared server. Each class is dynamically assigned a specific proportion of the server, based on the expected steady-state behaviour of a stationary queue with the same arrival and service rates (see Agnew (1976)).

3.1.5 Priority

Unlike traditional queueing systems where all customers are treated according to the same policies, multi-class systems can have different service disciplines for different classes of customers within a single queue. In multi-class systems, priority queueing disciplines are a common way to balance the requirements of different types of customers.

There are three main ways of implementing priority disciplines in multi-class queues: *preemptive priority*, *head of line priority* and *discretionary priority* (Jaiswal, 1968). In preemptive priority queues, the service of low priority customers is interrupted whenever a high priority customer joins the queue to allow the high priority customer to be serviced immediately. This does not occur in head of line priority queues, where high priority customers must wait in the queue until a server becomes available. Discretionary queues may apply a mixture of these rules depending on the state of the system, and may also allow for different classes to change priority.

The OPD queues follow a head of line priority discipline. This type of queueing system was introduced in Cobham (1954), which considers the equilibrium distribution of the waiting times in priority queueing systems. Most of the existing literature on this type of priority discipline deals with single-queue systems, rather than networks. For example, Bertsimas & Mourtzinou (1997) extends the priority model to customer classes with general arrival and service distributions in single-server queues with heavy traffic conditions. Huang *et al.* (2015) examines the role of priority systems in heavy-traffic healthcare queues and present priority strategies to deal with congestion in a single Emergency Department queue with feedback and deadlines.

Networks of priority queues were first treated by Morris (1981), where the joint probability distribution is derived for the number of customers from each class in the queues in a two-node, closed, Markovian network under equilibrium conditions. Fluid models for stationary priority networks are presented in Liu & Gong (2003) and priority networks with non-stationary arrivals are considered in Sharma & Tipper (1993).

Stellenbosch University https://scholar.sun.ac.za

The OPD fluid approximation models in § 3.4–§ 3.5 are similar to the method proposed in Sharma & Tipper (1993), although modifications have been made to incorporate the OPD's time dependent staff schedules and to allow the mean treatment times for each patient profile to differ. The models also use different profile weighting schemes to represent the server utilisation by each patient profile.

3.2 Variables and notation

This section introduces the general variables and notation that are used in all three queueing theory models. Calculations for the OPD routing matrix are also presented.

3.2.1 Queue length

All the models in this chapter are concerned with predicting the length and composition of each queue in the OPD over the course of the day. The variables $q_i(t)$ represent the total number of patients in the queue for each process $i \in \mathcal{I}$ at time t.

To understand the composition of each queue, the models also consider how many of its patients belong to each patient profile. Each queue is therefore divided into m sub-queues which represent the different patient profiles. The variables $q_i^p(t)$ indicate the number of patients from profile p queueing at process i at time t. The models in this chapter focus on the sub-queues, since the total queue length at each process can be calculated as $q_i(t) = \sum_{p \in \mathcal{P}} q_i^p(t)$.

3.2.2 Arrival routing probabilities

All the models in this chapter use routing probabilities to describe where a patient will go when they arrive at the OPD, and how they will move from process to process once they have entered the system.

In the conceptual model, the variables ϱ_i^p indicate the probability that a patient from profile p will visit process i during a visit to the OPD, irrespective of which other processes they visit. The order in which patients go to each process is specified by the variables ϕ_i^p, which rank the processes from first ($\phi_i^p = 1$) to last ($\phi_i^p = n$). The routing probabilities are calculated by combining these parameters to describe a patient's next move based on their current location in the system.

First, consider the routing of new arrivals from outside of the OPD network. The constants a_i^p indicate the probability that a patient from profile p will go directly to process i on arrival. Due to the assumption that no patients may leave the OPD without seeking treatment, the condition $\sum_{i \in \mathcal{I}} a_i^p = 1$ should hold for each profile p. Additionally, the routing probabilities should not allow patients go to processes that they do not require, so $a_i^p = 0$ whenever $\varrho_i^p = 0$. The constants a_i^p are calculated using the equation

$$a_i^p = a_*^p \varrho_i^p \prod_{j \in \mathcal{J}_i^p} (1 - \varrho_j^p), \qquad \mathcal{J}_i^p = \{j \mid j \in \mathcal{I} \text{ and } \phi_j^p < \phi_i^p\}, \tag{3.1}$$

where the set $\mathcal{J}_i^p \subset \mathcal{I}$ contains all processes that a patient from profile p might visit before they go to process i.

Stellenbosch University https://scholar.sun.ac.za

The constant a_*^p ensures that patients cannot leave the OPD without visiting at least one process. This condition is automatically satisfied for profiles that have a compulsory process (at least one $\varrho_i^p = 1$), and so $a_*^p = 1$ in these cases. If there are no compulsory processes, then a_*^p rescales the routing probabilities so that they add up to 1, i.e.

$$a_*^p = \left(\sum_{i \in \mathcal{I}} \left(\varrho_i^p \prod_{j \in \mathcal{J}_i^p} (1 - \varrho_j^p) \right) \right)^{-1}. \tag{3.2}$$

For convenience, the arrival routing probabilities are used to modify the arrival rate functions, $\eta_p \alpha^p(t)$, which give the rate at which new profile p patients arrive at the OPD in the conceptual model. These arrival rates are split into separate components for each process by multiplying these functions by the arrival routing probabilities. The process-specific arrival functions

$$\alpha_i^p(t) = a_i^p \eta_p \alpha^p(t), \qquad \text{with } i \in \mathcal{I},\ p \in \mathcal{P}, \tag{3.3}$$

indicate the arrival rate of new patients from profile p at process i.

3.2.3 Network routing matrix

Once patients have completed their first process they move through the network according to an $n \times n$ routing matrix. The routing matrix for patients from profile p is given by $\mathbf{R}^p$, where the element $r_{j,i}^p$ gives the probability that a patient who has just finished at process j will go directly to process i. The routing probabilities at each process do not necessarily add up to 1, since patients may leave the network instead of going to another process.

Since patients must visit processes in a specific order, $r_{j,i}^p = 0$ whenever process i is ranked before process j in the order variables (i.e. $\phi_i^p < \phi_j^p$), and $r_{j,i}^p = \varrho_i^p$ if process i follows directly after process j. The remaining elements of the routing matrix are calculated using a similar approach to the arrival probabilities, specifically

$$r_{j,i}^p = \varrho_i^p \prod_{k \in \mathcal{K}_{j,i}^p} (1 - \varrho_k^p), \qquad \text{when } \ \phi_j^p < \phi_i^p - 1. \tag{3.4}$$

In this case, the set $\mathcal{K}_{j,i}^p$ is a subset of $\mathcal{I}$ containing all processes that profile p patients might visit between process j and i, so

$$\mathcal{K}_{j,i}^p = \{k \mid k \in \mathcal{I} \ \text{ and } \ \phi_j^p < \phi_k^p < \phi_i^p\}. \tag{3.5}$$

In summary, the combined equation for the routing probabilities is

$$r_{j,i}^p = \begin{cases} 0, & \text{if } \phi_j^p \geq \phi_i^p, \\ \varrho_i^p, & \text{if } \phi_j^p + 1 = \phi_i^p, \\ \varrho_i^p \prod\limits_{k \in \mathcal{K}_{j,i}^p} (1 - \varrho_k^p), & \text{otherwise.} \end{cases} \tag{3.6}$$

3.3 Chapman-Kolmogorov equations

This section outlines a model of the OPD system as a continuous-time Markov process. The equations in this model give a probabilistic description of the OPD queues which could be used to

Stellenbosch University https://scholar.sun.ac.za

determine the expected length and variance of each queue, and approximate other performance measures such as expected waiting times.

This model is based on the Chapman-Kolmogorov equations,

$$\frac{\mathrm{d}\boldsymbol{\pi}(t)}{\mathrm{d}t} = \boldsymbol{\pi}(t)\mathbf{M}(t), \tag{3.7}$$

where $\boldsymbol{\pi}(t)$ is a vector of probabilities and $\mathbf{M}(t)$ is the transition rate matrix. The next two sections define the state space for this system (§ 3.3.1) and explain how the transition matrix is calculated (§ 3.3.2).

3.3.1 State space

The state of the OPD system at time t is given by the vector of random variables

$$\mathbf{y}(t) = (q_1^1, q_1^2, \ldots, q_1^m, \ldots, q_i^1, q_i^2, \ldots, q_i^m, \ldots, q_n^1, \ldots, q_n^m), \tag{3.8}$$

where q_i^p is the number of patients from profile $p \in \mathcal{P}$ in the queue at process $i \in \mathcal{I}$. The state space for $\mathbf{y}(t)$ is denoted by the set $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \ldots\}$. Since the variables q_i^p can only take non-negative integer values, the state space for this problem is discrete.

The state space for the OPD system would be infinite if there were no restrictions on the maximum length of each queue, but in the OPD conceptual model this is not the case. The conceptual model assumes that the number of patients that come to the OPD on a given day will lie within a certain range $\eta_p^{(l)} \leq \eta_p \leq \eta_p^{(u)}$, so $\mathbf{y}(t)$ can only assume states where the queue lengths are restricted by the condition $q_i^p \leq \eta_p^{(u)}$.

The size of the state space can be further reduced by imposing a stronger condition on the sub-queues for each patient profile,

$$\sum_{i \in \mathcal{I}} q_i^p \leq \eta_p^{(u)}. \tag{3.9}$$

This condition ensures that the total number of patients across all processes at any given time does not exceed $\eta_p^{(u)}$. However, it does not prevent the total daily arrivals from exceeding this limit, since it does not count patients who have already left the system before time t.

An important assumption of any Markov process is that the behaviour of the system at time t depends only on the current state $\mathbf{y}(t)$, and is not affected by anything that occurred before t. To limit the total daily arrivals for each profile to the upper bounds $\eta_p^{(u)}$, information about the cumulative number of arrivals that have occurred before time t must be included in the current state of the system.

This can be achieved by adding m pre-arrival queues to the state vectors, denoted by the variables $q_0^1, q_0^2, \ldots, q_0^m$. These queues indicate the number of patients from each profile who have not yet entered the system at time t, but may still do so. The new state vectors have the form

$$\mathbf{y}(t) = (q_0^1, \ldots, q_0^m, q_1^1, \ldots, q_i^p, \ldots, q_n^m). \tag{3.10}$$

At the beginning of each day, the OPD process queues are empty and all patients are in the pre-arrival queues. The initial state of the system is therefore

$$\mathbf{y}(0) = \left(q_0^1 = \eta_1^{(u)}, q_0^2 = \eta_2^{(u)}, \ldots, q_0^m = \eta_m^{(u)}, q_1^1 = 0, \ldots, q_i^p = 0, \ldots, q_n^m = 0\right). \tag{3.11}$$

Stellenbosch University https://scholar.sun.ac.za

By specifying that new patients can only enter the OPD through the pre-arrival queues, no further arrivals will occur for any profile $p \in \mathcal{P}$ once the system reaches a state where $q_0^p = 0$. The total number of arrivals will therefore never exceed the initial length of the pre-arrival queues, and each state in $\mathcal{X}$ is subject to the conditions

$$\left(q_0^p + \sum_{i \in \mathcal{I}} q_i^p \right) \leq \eta_p^{(u)}, \quad p \in \mathcal{P}. \tag{3.12}$$

3.3.2 Transition rates

The transition rate matrix $\mathbf{M}(t)$ in equation (3.7) describes how quickly the state of the system changes from the current state $\mathbf{y}(t) = \mathbf{x}_a$ to a different state $\mathbf{y}(t + \Delta t) = \mathbf{x}_b$. These transitions between states occur instantaneously at discrete points in time, since they are linked to discrete events.

The elements of $\mathbf{M}(t)$ are defined by the equations

$$M_{a,b}(t) = \begin{cases} \lim\limits_{\Delta t \to 0} \dfrac{\mathrm{P}(\mathbf{y}(t + \Delta t) = \mathbf{x}_b | \mathbf{y}(t) = \mathbf{x}_a)}{\Delta t}, & \text{if } a \neq b, \\ \lim\limits_{\Delta t \to 0} \dfrac{\mathrm{P}(\mathbf{y}(t + \Delta t) = \mathbf{x}_b | \mathbf{y}(t) = \mathbf{x}_a) - 1}{\Delta t}, & \text{if } a = b, \end{cases} \tag{3.13}$$

where $\mathrm{P}(\mathbf{y}(t + \Delta t) = \mathbf{x}_b | \mathbf{y}(t) = \mathbf{x}_a)$ is the probability that the system will be in state $\mathbf{x}_b$ at time $t + \Delta t$ given that the previous state was $\mathbf{x}_a$ at time t.

Transition probability assumptions

In this model, it is assumed that the inter-arrival and service times at each OPD queue are exponentially distributed. The probability that the time between two consecutive arrival or service events is less than s is therefore

$$F_t(s) = 1 - \exp\left(\frac{-s}{\theta(t)} \right), \tag{3.14}$$

where the parameter $\theta(t)$ represents the mean time between two consecutive events. To incorporate time-dependent components, this parameter may vary over the course of the day.

Substituting equation (3.14) into equation (3.13) gives the general formula for non-diagonal elements ($a \neq b$) in the transition rate matrix,

$$\begin{aligned} M_{a,b}(t) &= \lim_{\Delta t \to 0} \left(\frac{\mathrm{P}(\mathbf{y}(t + \Delta t) = \mathbf{x}_b | \mathbf{y}(t) = \mathbf{x}_a)}{\Delta t} \right) \\ &= \lim_{\Delta t \to 0} \left(\frac{1 - \exp\left(\dfrac{-\Delta t}{\theta(t)} \right)}{\Delta t} \right) \\ &= \lim_{\Delta t \to 0} \left(\frac{\dfrac{1}{\theta(t)} \exp\left(\dfrac{-\Delta t}{\theta(t)} \right)}{1} \right) \qquad \text{(L'Hôpital)} \\ &= \frac{1}{\theta(t)}. \end{aligned} \tag{3.15}$$

Stellenbosch University https://scholar.sun.ac.za

The corresponding formula for the diagonal elements of $\mathbf{M}(t)$ is therefore

$$\begin{aligned} M_{a,a}(t) &= \lim_{\Delta t \to 0} \left(\frac{\mathrm{P}(\mathbf{y}(t+\Delta t) = \mathbf{x}_a | \mathbf{y}(t) = \mathbf{x}_a) - 1}{\Delta t} \right) \\ &= \lim_{\Delta t \to 0} \left(\frac{1 - \sum_{b \neq a} \mathrm{P}(\mathbf{y}(t+\Delta t) = \mathbf{x}_b | \mathbf{y}(t) = \mathbf{x}_a) - 1}{\Delta t} \right) \\ &= \sum_{b \neq a} \lim_{\Delta t \to 0} \left(\frac{-\mathrm{P}(\mathbf{y}(t+\Delta t) = \mathbf{x}_b | \mathbf{y}(t) = \mathbf{x}_a)}{\Delta t} \right) \\ &= -\sum_{b \neq a} M_{a,b}. \end{aligned} \tag{3.16}$$

Transition rate calculations

The transition rate matrix for the OPD system can be calculated based on equations (3.15)–(3.16) and the parameters in the OPD conceptual model. This procedure is summarised below.

Step 1: Identify pairs of states $(\mathbf{x}_a, \mathbf{x}_b)$ where no transition is possible, and set $M_{a,b}(t) = 0$.

Step 2: Identify infeasible transitions $(\mathbf{x}_a \to \mathbf{x}_b)$ and set $M_{a,b}(t) = 0$.

Step 3: Calculate $M_{a,b}(t)$ for feasible non-diagonal transitions using equation (3.15).

Step 4: Calculate diagonal elements of $M_{a,a}(t)$ using equation (3.16).

Step 4 is relatively straightforward, but Steps 1, 2 and 3 require detailed knowledge of the OPD processes and patient profiles. Each of these steps is discussed in detail in the remainder of this section.

Step 1: Identify pairs of states $(\mathbf{x}_a, \mathbf{x}_b)$ where no transition is possible, and set $M_{a,b}(t) = 0$.

The transition matrix $\mathbf{M}(t)$ is very sparse, because the system can only transition from state $\mathbf{x}_a$ to $\mathbf{x}_b$ if the differences between these two states are the result of a single event. In this context, the types of events that cause the state of the system to change are the movements of individual patients.

When the system's current state is $\mathbf{y}(t) = \mathbf{x}_a = (q_0^1, \ldots, q_0^m, \ldots, q_i^p, \ldots, q_n^m)$, there are three types of events that may cause the system to transition out of this state:

Type 1: A new patient from profile p arrives and goes to process i:

$$(q_0^1, \ldots, q_0^p, \ldots, q_i^p, \ldots, q_n^m) \to (q_0^1, \ldots, q_0^p - 1, \ldots, q_i^p + 1, \ldots, q_n^m). \tag{3.17}$$

Type 2: A patient from profile p moves from process i to process j:

$$(q_1^0, \ldots, q_0^m, \ldots, q_i^p, \ldots, q_j^p, \ldots, q_n^m) \to (q_1^0, \ldots, q_0^m, \ldots, q_i^p - 1, \ldots, q_j^p + 1, \ldots, q_n^m). \tag{3.18}$$

Type 3: A patient from profile p finishes process i and leaves the OPD:

$$(q_0^1, \ldots, q_0^m, \ldots, q_i^p, \ldots, q_n^m) \to (q_1^0, \ldots, q_i^p - 1, \ldots, q_n^m). \tag{3.19}$$

Stellenbosch University https://scholar.sun.ac.za

Transitions $(\mathbf{x}_a \to \mathbf{x}_b)$ that do not match one of the three patterns in (3.17)–(3.19) cannot occur in a single event, so the corresponding transition rate is $M_{a,b}(t) = 0$.

Step 2: Identify infeasible transitions $(\mathbf{x}_a \to \mathbf{x}_b)$ and set $M_{a,b}(t) = 0$.

Step 2 eliminates infeasible transition pairs that are not identified in Step 1. Infeasible transitions match one of the patterns in (3.17)–(3.19), but the probability of these transitions occurring is 0 due to the configuration of the OPD system. The feasibility of transitions can be determined using the following criteria:

Type 1: Arrival event transitions of the form (3.17) are infeasible if a_i^p (the corresponding arrival routing probability) is 0.

Type 2: Inter-process transitions of the form (3.18) are infeasible if $r_{i,j}^p$ (the corresponding routing probability) is 0.

Type 3: Exit transitions of the form (3.19) are infeasible if the sum of the routing probabilities from the corresponding process i is 1, i.e. $\sum_{j \in \mathcal{I}} r_{i,j}^p = 1$.

After setting $M_{a,b}(t) = 0$ for all infeasible transitions, the remaining non-diagonal elements of $\mathbf{M}(t)$ represent events that do occur with some non-zero probability in the OPD model.

Step 3: Calculate $M_{a,b}(t)$ for feasible non-diagonal transitions using equation (3.15).

For a feasible transition $(\mathbf{x}_a \to \mathbf{x}_b)$, the corresponding transition rate is

$$M_{a,b}(t) = \begin{cases} \dfrac{1}{\theta(t)}, & \text{if } \theta(t) > 0, \\ 0, & \text{otherwise.} \end{cases} \tag{3.20}$$

It is necessary to use piecewise functions in equation (3.20) because the parameters $\theta(t)$ are not restricted to non-zero values.

Type 1: Arrival event transitions

For transitions associated with the arrival of a new patient, it is relatively simple to calculate the transition rates. These transition rates do not depend on the current state of the system, only the patient's profile $p \in \mathcal{P}$ and the patient's first process $i \in \mathcal{I}$. The rate at which new patients from profile p enter the system at process i is given by the function $\alpha_i^p(t)$ in equation (3.3), so the mean time between two arrival events is $\theta(t) = (\alpha_i^p(t))^{-1}$. The corresponding transition rates for feasible transitions of the form (3.17) are therefore

$$M_{a,b}(t) = \alpha_i^p(t). \tag{3.21}$$

Types 2–3: Treatment completion transitions

The transition rates for transitions of the form (3.18) and (3.19) are very similar, since both are related to patients completing treatment at their current process. This discussion focuses on cases where patients move from one queue to another queue within the network, and then generalises this method to include cases where patients exit the OPD.

Treatment completion events are more complicated than arrival events, because their transition rates depend on the current state, $\mathbf{x}_a$, as well as the patient's profile (p), which process they complete (i), and which queue they join next (j).

Stellenbosch University https://scholar.sun.ac.za

In the OPD conceptual model, the service time for profile p patients at process i is a random variable with a probability density function $f_{\tau_i^p}(x)$. In this model, it is assumed that these service times are exponentially distributed with mean $\bar{\tau}_i^p$, and that a single staff member at process i will treat patients from profile p at an average rate of $(\bar{\tau}_i^p)^{-1}$ patients per minute.

In periods when multiple staff are available at a particular process, these staff members function as independent servers with identical exponential service times. When more than one patient is treated at the same time, the average time between consecutive treatment completion events is reduced. The corresponding increase in the rate at which these events occur is directly proportional to the number of concurrent treatments in progress.

In simple cases where all patients at process i are from profile p and there are $\varsigma_i(t)$ staff on duty, the rate at which patients leave this process is $v_i^p(t)/\bar{\tau}_i^p$ patients per minute, where

$$v_i^p(t) = \min\left[q_i^p(t), \varsigma_i(t)\right]. \tag{3.22}$$

Since $v_i^p(t)$ depends on the queue length $q_i^p(t)$, the current state of the system influences the rate at which treatment completion events occur.

If there is more than one type of patient at a particular process, then the variables $v_i^p(t)$ reflect the number of staff that are treating patients from each of the different profiles. For convenience, let $p_{(1)}$ be the profile with the highest priority at process i. Assuming that staff will always treat higher priority patients first, the average treatment rate for patients in the sub-queue $q_i^{p_{(1)}}$ is $\frac{v_i^{p_{(1)}}(t)}{\bar{\tau}_i^{p_{(1)}}}$.

If there are fewer high priority patients than staff, i.e.

$$v_i^{p_{(1)}}(t) < \varsigma_i(t), \tag{3.23}$$

then the remaining staff will treat patients from the sub-queue with the second highest priority, $q_i^{p_{(2)}}$. The average treatment rate for this sub-queue is $v_i^{p_{(2)}}(t)/\bar{\tau}_i^{p_{(2)}}$ patients per minute, where

$$v_i^{p_{(2)}}(t) = \min\left[q_i^{p_{(2)}}(t),\ \varsigma_i(t) - v_i^{p_{(1)}}(t)\right]. \tag{3.24}$$

This calculation is repeated for each sub-queue in order of decreasing priority, assigning staff to each profile based on the formula

$$v_i^{p_{(k)}}(t) = \min\left[q_i^{p_{(k)}}(t),\ \varsigma_i(t) - \sum_{l=1}^{k-1} v_i^{p_{(l)}}(t)\right]. \tag{3.25}$$

This approach is sufficient to calculate the rate at which different types of patients are leaving process i when there is a clear hierarchy in which no two profiles have the same priority. However, most of the OPD queues do not work this way in reality, and multiple sub-queues often share the same priority.

To calculate the treatment rates for processes with priority ties, the patient profiles are divided into groups in descending order of priority. The highest priority group is denoted by the set

$$\mathcal{H}_i^{(1)} = \{h \in \mathcal{P} \mid \vartheta_i^h \leq \vartheta_i^p, \quad p \in \mathcal{P}\} \tag{3.26}$$

and the total number of staff assigned to these high priority sub-queues is

$$v_i^{(1)}(t) = \min\left[\sum_{p \in \mathcal{H}_i^{(1)}} q_i^p(t),\ \varsigma_i(t)\right]. \tag{3.27}$$

Stellenbosch University https://scholar.sun.ac.za

If the number of available staff is smaller than the total number of high priority patients, the variables $\beta_i^p(t)$ represent the expected proportion of the available staff that would be treating patients from each profile in $\mathcal{H}_i^{(1)}$. These proportions are calculated using the formula

$$\beta_i^p(t) = \frac{q_i^p \bar{\tau}_i^p}{\sum_{h \in \mathcal{H}_i^{(1)}} q_i^h \bar{\tau}_i^h} \times v_i^{(1)}(t), \quad p \in \mathcal{H}_i^{(1)}, \tag{3.28}$$

and the expected treatment rates for patients in these sub-queues is $\frac{\beta_i^p(t)}{\bar{\tau}_i^p}$.

This approach is not necessary if the number of available staff equals or exceeds the combined number of patients in the high priority queues. In these cases, exactly $q_i^h(t)$ staff members are assigned to each of the sub-queues in $\mathcal{H}_i^{(1)}$ and the corresponding treatment rates for patients in these queues are $\frac{q_i^h(t)}{\bar{\tau}_i^h}$.

The remaining staff are assigned to the sub-queues with the next highest priority,

$$\mathcal{H}_i^{(2)} = \{h \in \mathcal{P} \mid \vartheta_i^h \leq \vartheta_i^p,\ p \notin \mathcal{H}_i^{(1)}\}, \tag{3.29}$$

and the procedure is repeated for each successive group of sub-queues in order of decreasing priority. The general formula for the number of staff assigned to each priority group $\mathcal{H}_i^{(k)}$ is therefore

$$v_i^{(k)}(t) = \begin{cases} \min\left[\sum_{p \in \mathcal{H}_i^{(1)}} q_i^p(t),\ \varsigma_i(t)\right], & \text{if } k = 1, \\ \min\left[\sum_{p \in \mathcal{H}_i^{(k)}} q_i^p(t),\ \varsigma_i(t) - \sum_{l=1}^{k-1} v_i^{(l)}(t)\right], & \text{otherwise.} \end{cases} \tag{3.30}$$

The proportion of these staff assigned to each sub-queue in the set $\mathcal{H}_i^{(k)}$ is

$$\beta_i^p(t) = \begin{cases} 0, & \text{if } q_i^p(t) = 0, \\ q_i^p(t), & \text{if } p \in \mathcal{H}_i^{(k)} \text{ and } v_i^{(k)}(t) \geq \sum_{h \in \mathcal{H}_i^{(k)}} q_i^h(t), \\ v_i^{(k)}(t) \times \frac{q_i^p \bar{\tau}_i^p}{\sum_{h \in \mathcal{H}_i^{(k)}} q_i^h \bar{\tau}_i^h}, & \text{if } p \in \mathcal{H}_i^{(k)} \text{ and } v_i^{(k)}(t) < \sum_{h \in \mathcal{H}_i^{(k)}} q_i^h(t). \end{cases} \tag{3.31}$$

Based on equations (3.30) and (3.31), the rate at which patients from profile p are leaving process i is $\frac{\beta_i^p(t)}{\bar{\tau}_i^p}$. These results are combined with the routing probabilities to determine the rate at which patients from profile p are leaving process i and going to process j. The final transition rates for these events are calculated using the formula

$$M_{a,b}(t) = r_{i,j}^p \frac{\beta_i^p(t)}{\bar{\tau}_i^p}, \qquad i \neq j. \tag{3.32}$$

Transitions associated with patients leaving the OPD after process i are a simple extension of this formula, where the probability of a patient proceeding to a specific process j is replaced

Stellenbosch University https://scholar.sun.ac.za

with the probability that the patient does not go to any of the other processes in the network. The transition rates for these events are therefore

$$M_{a,b}(t) = \left(1 - \sum_{j \in \mathcal{I}} r_{i,j}^p\right) \frac{\beta_i^p(t)}{\bar{\tau}_i^p}. \tag{3.33}$$

Summary of the transition rate matrix

A summary of the different types of transition events and the corresponding transition rates is provided in Table 3.1.

Event Description	$\mathbf{x}_a \to \mathbf{x}_b$	Feasibility condition	$M_{a,b}(t)$
A new patient from profile p arrives and goes to process i.	$q_0^p \to q_0^p - 1$, $q_i^p \to q_i^p + 1$	$a_i^p > 0$	$\alpha_i^p(t)$
A patient from profile p moves from process i to process j.	$q_i^p \to q_i^p - 1$, $q_j^p \to q_j^p + 1$	$r_{i,j}^p > 0$	$r_{i,j}^p \frac{\beta_i^p(t)}{\bar{\tau}_i^p}$
A patient from profile p finishes process i and leaves the OPD.	$q_i^p \to q_i^p - 1$	$\sum_{j \in \mathcal{I}} r_{i,j}^p < 1$	$\left(1 - \sum_{j \in \mathcal{I}} r_{i,j}^p\right) \frac{\beta_i^p(t)}{\bar{\tau}_i^p}$

TABLE 3.1: *A summary of the transition rate matrix* $\mathbf{M}(t)$.

3.3.3 Solution

The solution to the Chapman-Kolmogorov equations is a set of functions

$$\pi_a(t) = \mathrm{P}\left(\mathbf{y}(t) = \mathbf{x}_a\right), \qquad \mathbf{x}_a \in \mathcal{X}, \tag{3.34}$$

which indicate the probability that the system $\mathbf{y}(t)$ will be in each of its possible states $\mathbf{x}_a \in \mathcal{X}$ at time t. Each of these functions is associated with a differential equation of the form

$$\frac{\mathrm{d}\pi_a(t)}{\mathrm{d}t} = \sum_{b=1}^{||\mathcal{X}||} \pi_b(t) M_{b,a}(t). \tag{3.35}$$

The equations can be simplified by removing the time dependence in $\mathbf{M}(t)$, which is due to the arrival functions $\alpha_i(t)$. These functions can be discretised and rewritten as piecewise constant functions. Since the staff schedules are already written as step functions, it is sensible to discretise the arrival rates over the same time intervals. For the purposes of this model, it is assumed that all staff changes occur at thirty minute intervals (i.e. 7h00, 7h30, 8h00,...). The new arrival functions are calculated as follows:

$$\varphi_i(t) = \int_{30(\lfloor t/30 \rfloor)}^{30(\lfloor t/30 \rfloor + 1)} \frac{\alpha_i(x)}{30} \mathrm{d}x \tag{3.36}$$

Using these functions, the transition rate matrix will be constant over each interval $[30(k-1), 30k]$ where $k \in \{1, 2, ..., 48\}$. The solution to the Chapman-Kolmogorov equations on each of these

Stellenbosch University https://scholar.sun.ac.za

intervals is then given by

$$\pi^k(t) = \pi^{k-1}(0)\ \mathrm{e}^{\mathbf{M}^{(k)}t},$$

where $\pi^0(0)$ is the initial state of the system. The final solution is a piecewise function

$$\pi(t) = \pi^{(k)}(t - 30k), \qquad \text{where } 30(k-1) \leq t < 30k. \tag{3.37}$$

Although the Chapman-Kolmogorov equations for the OPD system are theoretically interesting, this model cannot be practically implemented due to the size of the state space. For example, an OPD set-up with 6 processes and 6 patient profiles has approximately 2.6×10^{35} different possible states. In a general system of n processes and m different patient profiles, the total size of the state space is

$$||\mathcal{X}|| = \prod_{p=1}^{m} \left(\sum_{k=K_p}^{n+1} \binom{n+1}{k} \times \binom{\eta_p^{(u)} + 1}{n+1-k} \right), \tag{3.38}$$

where $K_p = \max[0, n - \eta_p^{(u)}]$.

The number of different combinations of queue lengths makes it impractical to enumerate and model all the possible states of the OPD system, so the rest of this chapter takes a slightly different approach to this problem. Sections 3.4 and 3.5 describe how the queue length variables can be modelled directly, without considering every possible state of the entire system.

3.4 Priority fluid model

This section describes a fluid approximation model of the OPD queue network. Unlike traditional queueing models, fluid approximations do not provide any information about the probability of the system being in a given state. Instead, fluid approximations for queueing networks model the expected length of each queue based on the mean flow of jobs through each node.

The assumptions of this approach are explained in § 3.4.1, along with an intuitive derivation of the fluid equations for the flow of patients into and out of each queue. These equations are reformulated in § 3.4.2 to reduce the number of variables in the system, and the numerical method used to solve the reduced system is discussed in § 3.4.3.

3.4.1 Model derivation

In this model the lengths of the OPD queues are approximated by a set of continuous, non-negative variables, $\{q_1(t), q_2(t), \ldots, q_n(t)\}$. The behaviour of these queues over the course of the day is determined by first-order differential equations

$$\frac{\mathrm{d}q_i(t)}{\mathrm{d}t} = \lambda_i(t) - \mu_i(t), \qquad \text{with } i \in \mathcal{I}, \tag{3.39}$$

where $\lambda_i(t)$ is the rate at which patients join queue i and $\mu_i(t)$ is the rate at which patients leave the queue. The variables $\lambda_i(t)$ and $\mu_i(t)$ do not require any assumptions about the arrival and service distributions, since they represent the mean arrival and service rates.

To account for the different patient profiles, each queue is modelled as the sum of m sub-queues, so that

$$q_i(t) = q_i^1(t) + \cdots + q_i^m(t), \qquad \text{with } i \in \mathcal{I}. \tag{3.40}$$

Stellenbosch University https://scholar.sun.ac.za

As in the previous model, the variables $q_i^p(t)$ give the number of patients from profile $p \in \mathcal{P}$ that are queueing at process $i \in \mathcal{I}$. The arrival and service rates can also be written in terms of separate components for each sub-queue, where

$$\lambda_i(t) = \lambda_i^1(t) + \cdots + \lambda_i^m(t), \qquad \text{with } i \in \mathcal{I}, \text{ and} \tag{3.41}$$

$$\mu_i(t) = \mu_i^1(t) + \cdots + \mu_i^m(t), \qquad \text{with } i \in \mathcal{I}. \tag{3.42}$$

Equation (3.40) is reformulated in terms of the sub-queues, resulting in a system of $n \times m$ differential equations,

$$\frac{\mathrm{d}q_i^p(t)}{\mathrm{d}t} = \lambda_i^p(t) - \mu_i^p(t), \qquad \text{with } i \in \mathcal{I} \text{ and } p \in \mathcal{P}. \tag{3.43}$$

The solution to the system in equation (3.43) approximates the expected length of each of the OPD sub-queues over the course of the day. The remainder of this section explains the arrival rate functions, $\lambda_i^p(t)$ and the service rate functions, $\mu_i^p(t)$.

Arrival rates

The variables $\lambda_i^p(t)$ are often called *effective arrival rates* because they represent the average rate at which patients from profile p are joining the queue at process i at time t. In open queue networks, the effective arrivals are a combination of new patients coming into the network (*external arrivals*), as well as patients moving from one process to another (*internal arrivals*).

The external arrival rates are given by the functions

$$\alpha_i^p(t) = a_i^p \eta_p \alpha^p(t), \qquad \text{with } i \in \mathcal{I} \text{ and } p \in \mathcal{P}, \tag{3.44}$$

which were introduced in § 3.2.2. The internal arrival rates are the sum of patients arriving at process i from any of the other OPD processes,

$$\sum_{j \in \mathcal{I}} r_{j,i}^p \mu_j^p(t), \tag{3.45}$$

where $\mathbf{R}^p$ is the $n \times n$ routing matrix calculated in § 3.2.3.

The effective arrival rate for patients from profile p at process i is the sum of the internal and external arrival rates in equations (3.44) and (3.45), and so

$$\lambda_i(t) = \alpha_i^p(t) + \sum_{j \in \mathcal{I}} r_{j,i}^p \mu_j^p(t), \qquad \text{with } i \in \mathcal{I} \text{ and } p \in \mathcal{P}. \tag{3.46}$$

Service rates

The service rates $\mu_i^p(t)$ are more challenging to represent as a fluid approximation because they are dependent on the number of staff on duty as well as the size of each sub-queue. For the purposes of this model, the staff at each process are treated as a single server with a variable service capacity, namely $\varsigma_i(t)$. When $\varsigma_i(t) > 0$, the average treatment rates increase linearly for each additional staff member.

The traffic intensity at each node in the fluid network is the ratio of incoming work to server capacity,

$$\rho_i(t) = \frac{1}{\varsigma_i(t)} \sum_{p \in \mathcal{P}} \bar{\tau}_i^p \lambda_i^p(t), \qquad \text{with } i \in \mathcal{I}. \tag{3.47}$$

Stellenbosch University https://scholar.sun.ac.za

At any given time, each process in the OPD will be either *underloaded* or *overloaded*. When a particular node is underloaded, there is no queue at this node and the traffic intensity is less than 1. If either of these conditions does not hold, the node is said to be overloaded.

In the fluid model, the service rates differ during underloaded and overloaded intervals. For convenience, the binary functions $\phi_i(t)$ are used to indicate whether each process is underloaded or overloaded, i.e.

$$\phi_i(t) = 1 \qquad \text{when} \qquad \sum_{p\in\mathcal{P}} \bar{\tau}_i^p \lambda_i^p(t) \leq \varsigma_i(t) \quad \text{and} \quad q_i(t) = 0, \tag{3.48}$$

$$\phi_i(t) = 0 \qquad \text{when} \qquad \sum_{p\in\mathcal{P}} \bar{\tau}_i^p \lambda_i^p(t) > \varsigma_i(t) \quad \text{or} \quad q_i(t) > 0. \tag{3.49}$$

When a node is underloaded, the length of each sub-queue is zero and there are enough staff on duty to treat each patient as they arrive, so when $\phi_i(t) = 1$, the service rates at process i are

$$\mu_i^p(t) = \lambda_i^p(t), \qquad \text{with } p \in \mathcal{P}. \tag{3.50}$$

When a node is overloaded, staff members must divide their time between the different sub-queues in order to ensure that every patient will eventually receive treatment. The variables β_i^p are used to represent the fraction of the total staff capacity at process i that is used to treat patients from profile p. The service rate functions for overloaded intervals are therefore

$$\mu_i^p(t) = \frac{\beta_i^p \varsigma_i(t)}{\bar{\tau}_i^p}, \qquad \text{with } i \in \mathcal{I} \text{ and } p \in \mathcal{P}. \tag{3.51}$$

By combining equations (3.48)–(3.52), the service rate functions can be summarised as

$$\mu_i^p(t) = \phi_i(t)\lambda_i^p(t) + (1 - \phi_i(t))\frac{\beta_i^p(t)\varsigma_i(t)}{\bar{\tau}_i^p}, \tag{3.52}$$

where

$$\phi_i(t) = \begin{cases} 1, & \text{if } \sum\limits_{p\in\mathcal{P}} \bar{\tau}_i^p \lambda_i^p(t) \leq \varsigma_i(t) \text{ and } \sum\limits_{p\in\mathcal{P}} q_i^p(t) = 0, \\ 0, & \text{otherwise.} \end{cases} \tag{3.53}$$

Profile weighting models

This section explains how the weights β_i^p are derived in the fluid approximation model. These weights are a continuous approximation of the β_i^p coefficients in the Markov model. However, the fluid approximation weights are only required when a particular queue is overloaded, whereas the β_i^p coefficients in the Markov model apply to both underloaded and overloaded queues. The fluid model weights are simpler than the coefficients in the discrete system, since they need not capture individual staff members or patients.

In the fluid model, the most important properties of these weights are as follows:

Property 1: The sum of the weights for any given process must add up to 1 during overloaded periods, i.e. $\sum\limits_{p\in\mathcal{P}} \beta_i^p = 1$.

Property 2: A weight of $\beta_i^p = 0$ must be assigned whenever there are no patients of type p at process i.

Stellenbosch University https://scholar.sun.ac.za

Property 3: Weights should be proportional to both the number of patients in the queue as well as the amount of time needed to treat each patient.

Property 4: Weights should be proportional to the relative priority of different patient profiles.

Properties 2 and 3 indicate that β_i^p must be a time dependent function, since fixed weights would not be able to reflect any changes in the composition of the queue. The time dependent weights have the form

$$\beta_i^p(t) \propto c_i(t) q_i^p(t) \bar{\tau}_i^p, \tag{3.54}$$

where $c_i(t)$ are scaling coefficients to ensure that the weights at each process add up to 1. The weights are divided by the priority parameters to incorporate the priority of each patient profile, and so

$$\beta_i^p(t) = \frac{c_i(t) q_i^p(t) \bar{\tau}_i^p}{\vartheta_i^p}. \tag{3.55}$$

These weights already satisfy Properties 2, 3 and 4. To satisfy Property 1,

$$\sum_{p \in \mathcal{P}} \frac{c_i(t) q_i^p(t) \bar{\tau}_i^p}{\vartheta_i^p} = 1, \qquad \text{and thus} \qquad c_i(t) = \left(\sum_{p \in \mathcal{P}} \frac{q_i^p(t) \bar{\tau}_i^p}{\vartheta_i^p} \right)^{-1}. \tag{3.56}$$

To avoid $c_i(t) = 0^{-1}$ when there is no queue, $\phi_i(t)$ is added to the terms in equation (3.56). This modification does not actually influence the weights, since they only apply during overloaded periods where $\phi_i(t) = 0$. The final equations for the profile weights are then

$$\beta_i^p(t) = \frac{q_i^p(t) \bar{\tau}_i^p}{\left(\sum_{k \in \mathcal{P}} \frac{q_i^k(t) \bar{\tau}_i^k}{\vartheta_i^k} + \phi_i(t) \right)}, \qquad \text{with } i \in \mathcal{I} \text{ and } p \in \mathcal{P}. \tag{3.57}$$

This implementation of the priority queueing discipline has slightly different implications to the priority queueing discipline in the Markov model. In the Markov model, it is not possible for a staff member to treat a low priority patient if there are any high priority patients in the queue who are not receiving treatment. Therefore, the treatment of low priority patients will be interrupted when a high priority patient joins the queue.

The weights in equation (3.55) represent a less strict priority system, where low priority patients may still receive treatment even if there are high priority patients in the queue. This is more likely to occur if the low priority sub-queues are much longer than the high priority sub-queues, or if the low priority patients have longer average treatment times. These weights are a better representation of the priority discipline in the OPD queues, since staff members do not generally interrupt a low priority patient's treatment to deal with a higher priority patient unless the other patient is exceptionally urgent.

Stellenbosch University https://scholar.sun.ac.za

3.4.2 Reformulation

The priority fluid model described in § 3.4.1 can be summarised in terms of five sets of equations, namely

$$q_i(t) = \sum_{p\in\mathcal{P}} q_i^p(t), \tag{3.58}$$

$$\frac{\mathrm{d}q_i^p(t)}{\mathrm{d}t} = \lambda_i^p(t) - \mu_i^p(t), \tag{3.59}$$

$$\lambda_i^p(t) = \alpha_i^p(t) + \sum_{j\in\mathcal{I}} r_{j,i}^p \mu_j^p(t), \tag{3.60}$$

$$\mu_i^p(t) = \phi_i(t)\lambda_i^p(t) + (1-\phi_i(t))\,\varsigma_i(t) q_i^p(t) \left(\sum_{k\in\mathcal{P}} \frac{q_i^k(t)\bar{\tau}_i^k}{\vartheta_i^k} + \phi_i(t)\right)^{-1} \tag{3.61}$$

$$\phi_i(t) = \begin{cases} 1, & \text{if } \sum\limits_{p\in\mathcal{P}} \lambda_i^p(t)\bar{\tau}_i^p \le \varsigma_i(t) \quad \text{and} \quad q_i(t) = 0, \\ 0, & \text{otherwise.} \end{cases} \tag{3.62}$$

This system consists of $3nm+2n$ different equations, which correspond with $3nm+2n$ unknown functions. In this section, unnecessary variables are eliminated from equations (3.58)–(3.62) to reduce the size of the system to only $nm+n$ equations.

First, the $q_i(t)$ variables are eliminated from equations (3.59)–(3.62) by replacing these variables with the sub-queue summations from equation (3.58). The length of each queue is fully described by the corresponding sub-queue variables, so equation (3.58) can be removed from the system.

In the remaining equations, the variables $\lambda_i^p(t)$ and $\mu_i^p(t)$ are defined by algebraic equations, while $q_i^p(t)$ is expressed in terms of a first-order differential equation. Since $\lambda_i^p(t)$ and $\mu_i^p(t)$ are rates of change in the length of each queue, it is more appropriate to treat these variables as first-order derivatives. New variables $\Lambda_i^p(t)$ and $U_i^p(t)$ are defined, where

$$U_i^p(t) = \int_0^t \mu_i^p(x)\mathrm{d}x, \qquad \text{and} \qquad \frac{\mathrm{d}U_i^p(t)}{\mathrm{d}t} = \mu_i^p(t), \tag{3.63}$$

$$\Lambda_i^p(t) = \int_0^t \lambda_i^p(x)\mathrm{d}x, \qquad \text{and} \qquad \frac{\mathrm{d}\Lambda_i^p(t)}{\mathrm{d}t} = \lambda_i^p(t). \tag{3.64}$$

In terms of the OPD queues, the functions $\Lambda_i^p(t)$ express the cumulative number of patients from profile p who have arrived at process i by time t, and $U_i^p(t)$ indicates the cumulative number of patients from profile p who have received treatment at process i by time t.

Stellenbosch University https://scholar.sun.ac.za

Substituting these variables into equations (3.59)-(3.62) gives the new system

$$\frac{\mathrm{d}q_i^p(t)}{\mathrm{d}t} = \frac{\mathrm{d}\Lambda_i^p(t)}{\mathrm{d}t} - \frac{\mathrm{d}U_i^p(t)}{\mathrm{d}t}, \tag{3.65}$$

$$\frac{\mathrm{d}\Lambda_i^p(t)}{\mathrm{d}t} = \alpha_i^p(t) + \sum_{j\in\mathcal{I}} r_{j,i}^p \frac{\mathrm{d}U_j^p(t)}{\mathrm{d}t}, \tag{3.66}$$

$$\frac{\mathrm{d}U_i^p(t)}{\mathrm{d}t} = \phi_i(t)\left(\alpha_i^p(t) + \sum_{j\in\mathcal{I}} r_{j,i}^p \frac{\mathrm{d}U_j^p(t)}{\mathrm{d}t}\right) + (1-\phi_i(t))\,\varsigma_i(t) q_i^p(t)\left(\sum_{k\in\mathcal{P}} \frac{q_i^k(t)\bar{\tau}_i^k}{\vartheta_i^k} + \phi_i(t)\right)^{-1}, \tag{3.67}$$

$$\phi_i(t) = \begin{cases} 1 & \text{if } \sum_{p\in\mathcal{P}} \frac{\mathrm{d}\Lambda_i^p(t)}{\mathrm{d}t}\bar{\tau}_i^p \le \varsigma_i(t) \quad \text{and} \quad q_i(t) = 0, \\ 0 & \text{otherwise.} \end{cases} \tag{3.68}$$

Equations (3.65) and (3.66) are linear first-order differential equations which can be integrated to find the solutions

$$q_i^p(t) = \Lambda_i^p(t) - U_i^p(t), \tag{3.69}$$

$$\Lambda_i^p(t) = A_i^p(t) + \sum_{j\in\mathcal{I}} r_{j,i}^p U_j^p(t). \tag{3.70}$$

The function $A_i^p(t)$ in equation (3.70) is the integral of the external arrival rate functions,

$$A_i^p(t) = \int_0^t \alpha_i^p(x)\mathrm{d}x. \tag{3.71}$$

By substituting equation (3.70) back into equation (3.69), the queue functions can be expressed in terms of only the $U_i^p(t)$ functions, where

$$q_i^p(t) = A_i^p(t) + \sum_{j\in\mathcal{I}} r_{j,i}^p U_j^p(t) - U_i^p(t). \tag{3.72}$$

At this point, the only remaining unknown functions in the system are $U_i^p(t)$ and $\phi_i(t)$, which are the solutions to equations (3.67) and (3.68). These equations are re-written in terms of only $U_i^p(t)$ and $\phi_i(t)$ by using equations (3.70) and (3.72) to eliminate all instances of $\Lambda_i^p(t)$ and $q_i^p(t)$.

Stellenbosch University https://scholar.sun.ac.za

This leads to the reduced system

$$\frac{\mathrm{d}U_i^p(t)}{\mathrm{d}t} = \phi_i(t)\left(\alpha_i^p(t) + \sum_{j\in\mathcal{I}} r_{j,i}^p \frac{\mathrm{d}U_j^p(t)}{\mathrm{d}t}\right) + (1-\phi_i(t))\left(\frac{\varsigma_i(t)\left(A_i^p(t) + \sum_{j\in\mathcal{I}} r_{j,i}^p U_j^p(t) - U_i^p(t)\right)}{\sum_{k\in\mathcal{P}} \frac{\bar{\tau}_i^k}{\vartheta_i^k}\left(A_i^k(t) + \sum_{j\in\mathcal{I}} r_{j,i}^k U_j^k(t) - U_i^k(t)\right) + \phi_i(t)}\right), \tag{3.73}$$

$$\phi_i(t) = \begin{cases} 1, & \text{if } \sum_{p\in\mathcal{P}}\left(\alpha_i^p(t) + \sum_{j\in\mathcal{I}} r_{j,i}^p \frac{\mathrm{d}U_j^p(t)}{\mathrm{d}t}\right)\bar{\tau}_i^p \le \varsigma_i(t) \quad \text{and} \\ & A_i^p(t) + \sum_{j\in\mathcal{I}} r_{j,i}^p U_j^p(t) - U_i^p(t) = 0, \\ 0, & \text{otherwise.} \end{cases} \tag{3.74}$$

The reduced system contains both differential and algebraic equations. These equations cannot be solved analytically, but the elimination of $\Lambda_i^p(t)$ and $q_i^p(t)$ simplifies the system enough to find numerical solutions for the remaining unknown functions.

3.4.3 Solution

Equations (3.73)–(3.74) were implemented in Mathematica (Wolfram Research, Inc.) using the *NDSolve* function. This function can be adapted to solve systems of differential equations with discrete variables and discontinuities. Documentation for this function is available online (Wolfram Research, Inc., 2014), and the general procedure for solving these systems is summarised below.

Numerical solution method for the PF model equations

Step 1: Initialise $t = 0$, $U_i^p(0) = 0$ and $\phi_i(t) = 0$

Step 2: Calculate all $U_i^p(t + \Delta t)$

Step 3: Calculate all $\phi_i(t + \Delta t)$

Step 4: If $\phi_i(t + \Delta t) \neq \phi_i(t)$:

- **4.1:** Locate discontinuity at $t + \delta$
- **4.2:** Calculate $U_i^p(t + \delta)$
- **4.3:** Calculate $\phi_i(t + \delta^+)$
- **4.4:** Update $t = t + \delta$, go to Step 6

Step 5: Update $t = t + \Delta t$

Step 6: If $t < t_{\text{final}}$, return to Step 2

Stellenbosch University https://scholar.sun.ac.za

3.5 FCFS fluid model

In this section, a second fluid approximation model for the OPD system is introduced. The key difference between this model and the previous models in § 3.3 and § 3.4 is that the queues in this model operate according to a strict first-come, first-serve discipline.

The model presented in this section is also based on a continuous approximation of the OPD system, but in this case the fluid analogy is not particularly useful. Instead, the flow of patients though each queue can be compared to particles in a grain silo, where the arrival rates $\lambda_i^p(t)$ represent the rate at which new grain is poured into the top of the silo and the service rates $\mu_i^p(t)$ represent the grain that is being taken out of the bottom of the silo. If the silo is supplied by m different farms, $p = 1, 2, \ldots, m$, then the composition of the grain in the silo will vary according to when each of these farms harvests their crop. Grain harvested earlier in the season will be closer to the bottom of the silo, so it will leave the silo first.

This approach is motivated by the fact that patient profiles in the OPD model can have very different arrival patterns. For example, patients who return to the OPD on a regular basis tend to plan their trips so that they arrive early in the day, while casualty arrivals are more evenly spread. These discrepancies influence the distribution of patients from different profiles within the queues. On busy days, there is usually a long queue of patients waiting to see the doctor by mid-morning. Most of the patients near the front of this queue will be returning patients, and there will be more casualty patients closer to the back of the queue.

This phenomenon is not represented in either of the models in § 3.3–§ 3.4. In these models, it is assumed that the rate at which different profiles are being treated is determined by **(a)** which patients have the highest priority; **(b)** which patients take the longest to treat; and **(c)** which patients make up the biggest portion of the total queue length. Neither of these models accounts for the fact that patients who have been waiting for the longest are more likely to be treated first.

3.5.1 Model derivation

The FCFS fluid model is based on the same system of equations that were introduced for the priority fluid model in § 3.4.1, namely

$$\frac{\mathrm{d}q_i^p(t)}{\mathrm{d}t} = \lambda_i^p(t) - \mu_i^p(t), \tag{3.75}$$

$$\lambda_i^p(t) = \alpha_i^p(t) + \sum_{j \in \mathcal{I}} r_{j,i}^p \mu_j^p(t), \tag{3.76}$$

$$\mu_i^p(t) = \phi_i(t)\lambda_i^p(t) + (1 - \phi_i(t))\frac{\varsigma_i(t)\beta_i^p(t)}{\bar{\tau}_i^p}, \tag{3.77}$$

$$\phi_i(t) = \begin{cases} 1, & \text{if } \sum_{p \in \mathcal{P}} \lambda_i^p(t)\bar{\tau}_i^p \leq \varsigma_i(t) \text{ and } q_i(t) = 0, \\ 0, & \text{otherwise.} \end{cases} \tag{3.78}$$

The weights $\beta_i^p(t)$ are adjusted to reflect the composition of different patient profiles at the front of each queue. When these weights are used, the functions $\mu_i^p(t)$ represent the rate at which patients are starting treatments at process i, rather than leaving the process. Equation (3.76) is therefore adjusted to allow a delay of $\bar{\tau}_i^p$ minutes between the start of treatments at process

Stellenbosch University https://scholar.sun.ac.za

i and the arrivals at another process in the network. The new arrival rate equations are

$$\lambda_i^p(t) = \alpha_i^p(t) + \sum_{j\in\mathcal{I}} r_{j,i}^p \; \mu_j^p\Big(t - \bar{\tau}_j^p\Big), \qquad \text{with } i \in \mathcal{I} \text{ and } p \in \mathcal{P}. \tag{3.79}$$

The rest of this section explains how the $\beta_i^p(t)$ weights for FCFS queues are calculated. The procedure can be summarised in the following two steps:

Step 1: Determine the arrival time of the patients who have been waiting in the queue for the longest period of time.

Step 2: Distribute staff according to the composition of the arrival rates when these patients arrived.

To avoid confusion, subscripts will be added to the independent variables in this model to differentiate between the *current time*, t, the *time of arrival* at a particular queue, t_a, and the *time of departure* from a queue, t_b. Using this notation, the weights $\beta_i^p(t_b)$ will reflect the composition of different patient profiles in the arrival rate functions $\lambda_i^p(t_a)$, i.e.

$$\beta_i^p(t_b) = \frac{\lambda_i^p(t_a)}{\sum\limits_{k\in\mathcal{P}} \lambda_i^k(t_a)}. \tag{3.80}$$

The relationship between the arguments t_a and t_b is described by a function $h_i(t_b) = t_a$, which maps the departure time of each patient to their arrival time. When the solutions for $\lambda_i(t)$ and μ_i are known, the function $h_i(t_b)$ satisfies the equations

$$h_i(t_b) = \Lambda_i^{-1}(U_i(t_b)), \qquad \text{with } i \in \mathcal{I}. \tag{3.81}$$

In this case, however, $\mu_i(t)$ are unknown functions which are dependent on $h_i(t_b)$, and so the $h_i(t_b)$ functions must be derived without knowing the service rates for each queue.

If the queues follow a FCFS policy, then $h_i(t_b)$ is a strictly increasing function on the domain $\{t_b : \varsigma_i(t_b) > 0\}$, and maps each value t_b to a unique point in the domain $\{t_a : \lambda_i(t_a) > 0\}$. Therefore, there is an invertible function $g_i(t_a) = t_b$ such that $h_i(t_b) = g_i^{-1}(t_b)$. The function $g_i(t_a)$ gives the time of treatment for a patient who arrives at process i at time t_a. Therefore,

$$g_i(t_a) = t_b = t_a + w_i(t_a), \tag{3.82}$$

where $w_i(t_a)$ is the waiting time delay between a patient's arrival time, t_a and the corresponding departure time t_b.

Since service times are deterministic in the fluid approximation models, the waiting time variables can be calculated from the length of each queue at the time of arrival. The first step in calculating $w_i(t_a)$ is to add up the total time needed to treat all the patients who are currently in the queue at time t_a, which is equal to the summation

$$\sum_{p\in\mathcal{P}} \bar{\tau}_i^p q_i^p(t_a). \tag{3.83}$$

The next step is to determine how long it will take for the staff members on duty to complete all of these treatments. With constant staff schedules, this can be achieved by dividing the total time by the number of staff.

Stellenbosch University https://scholar.sun.ac.za

Complications arise when the number of staff changes over the course of a day. The effect of any staff changes that will occur while that patient is still waiting in the queue must be considered, since they will affect the patient's waiting time. As in the Markov model, it is assumed that all staff changes occur at thirty minute intervals.

The cumulative number of staff hours for all staff on duty at process i up until time t is given by the piecewise function

$$S_i(t) = 30\left(\sum_{k=0}^{\lfloor t/30 \rfloor} \varsigma_i(30k)\right) - \varsigma_i(t)\text{mod}[k, 30]. \tag{3.84}$$

The change in this function over an interval $[t_1, t_2]$ represents the total amount of treatment time available at process i during that period.

If a patient arrives at time t_a and leaves at t_b, then the total amount of treatment time available in this interval must be equal to the total amount of time required to treat all patients who were in the queue at time t_a. Therefore,

$$S_i(t_b) - S_i(t_a) = \sum_{p\in\mathcal{P}} \bar{\tau}_i^p q_i^p(t_a). \tag{3.85}$$

Replacing the departure time t_b in equation (3.85) with $t_a + w_i(t_a)$ gives

$$S_i(t_a + w_i(t_a)) = \sum_{p\in\mathcal{P}} \bar{\tau}_i^p q_i^p(t_a) + S_i(t_a), \tag{3.86}$$

which can be used to calculate $w_i(t_a)$ in terms of t_a.

The function $S_i(t_a + w_i(t_a))$ must be inverted to solve equation (3.86) for $w_i(t_a)$. Since $S_i(t)$ is strictly increasing whenever there is at least one staff member on duty, an inverse function $S_i^{-1}(t)$ can be defined on the domain $\{t : \varsigma_i(t) > 0\}$, such that

$$S_i^{-1}(t) = \begin{cases} 30\left(k + \dfrac{S_i(30k) - t}{S_i(30k) - S_i(30(k-1))}\right), & \text{if } S_i(30(k-1)) \le t < S_i(30k), \\ 30k, & \text{if } S_i(30(k-1)) = t = S_i(30k). \end{cases} \tag{3.87}$$

with $k \in \{1, \ldots, 48\}$.

Applying the inverse function to both sides of equation (3.86) gives

$$w_i(t_a) = S_i^{-1}\left(\sum_{p\in\mathcal{P}} \bar{\tau}_i^p q_i^p(t_a) + S_i(t_a)\right) - t_a, \tag{3.88}$$

and substituting this into equation (3.82) results in

$$g_i(t_a) = S_i^{-1}\left(\sum_{p\in\mathcal{P}} \bar{\tau}_i^p q_i^p(t_a) + S_i(t_a)\right). \tag{3.89}$$

The final step in this process is to calculate $h_i(t_b)$ by inverting $g_i(t_a)$. This step can be very time-consuming when $g_i(t_a)$ contains many piecewise components that must all be treated individually, especially if the function is non-linear on these intervals.

To simplify these calculations, the external arrival rate functions $\alpha_i^p(t)$ are discretised as piecewise constant functions over the same thirty minute intervals as the staff schedules. This adjustment minimises the number of discontinuities in $g_i(t_a)$ and ensures that the functions $\mu_i^p(t)$ and $\lambda_i^p(t)$ are piecewise constant, while $q_i^p(t)$, $g_i(t_a)$ and $h_i(t_b)$ are piecewise linear functions.

Stellenbosch University https://scholar.sun.ac.za

3.5.2 Solution

Unlike the priority fluid model, most of the equations in the FCFS model can be solved analytically. The only numerical calculations required are the switching points in the functions $\phi_i(t)$, which indicate the beginning or end of an overloaded period. The procedure for solving the equations for each process is summarised below.

Solving the FCFS fluid equations

Step 1: Initialise $t^* = 0$

Step 2: Determine the arrival rate functions $\lambda_i^p(t)$

Step 3: Calculate the traffic intensity $\rho_i(t)$

Step 4: Set $\mu_i^p(t) = \lambda_i^p(t)$

Step 5: Set $q_i^p(t) = 0$

Step 6: If there is a switching point after t^* where $\phi_i(s_1) = 0$:

- **6.1** Set $\mu_i^p(t) = \lambda_i^p(t)$ for $t^* \leq t \leq s_1$.
- **6.2** Calculate $\beta_i^p(t)$ for $s_1 < t$.
- **6.3** Set $\mu_i^p(t) = \dfrac{\varsigma_i(t)\beta_i^p(t)}{\bar{\tau}_i^p}$ for $s_1 < t$
- **6.4** Set $q_i^p(t) = \int_{s1}^{t} \lambda_i^p(x) - \mu_i^p(x)\mathrm{d}x$ for $s_1 < t$

Step 7: If there is a switching point after t^* where $\phi_i(s_2) = 1$:

- **7.1** Set $\mu_i^p(t) = \lambda_i^p(t)$ for $s_2 \leq t$.
- **7.2** Set $q_i^p(t) = 0$ for $s_2 \leq t$
- **7.3** Update $t^* = s_2$
- **7.4** Return to Step 6

This procedure is relatively straightforward in systems where all the routing matrices $\mathbf{R}^p$ are upper triangular matrices, since the equations for each of the different processes can be solved separately. The first process $i = 1$ will have no internal arrivals, so this solution is calculated first. The remaining processes are then solved in order, using the service rate solutions that have already been found for $\mu_1^p(t), \ldots, \mu_i^p(t)$ to calculate the arrival rates at process $i + 1$.

Routing matrices that are not upper triangular matrices indicate that there is a feedback loop between two or more processes in the network. This occurs when the order of some processes is reversed for different patient profiles, so there are patients moving from queue i to queue j and also from queue j back to queue i. In this case, the solution for queue i is needed to calculate the solution for queue j, but the solution for queue i also depends on queue j.

Systems with routing loops can be solved iteratively over a series of time intervals. The iterative process exploits the delay of $\bar{\tau}_i^p$ minutes between a patient starting treatment at process i and arriving at any other process in the network. Each iteration calculates partial solutions for the arrival rates and service rates at each process, up to the point that the required internal arrival rate solutions are known. Using this method, the partial solutions for each process are iteratively extended until the complete solutions are known.

Stellenbosch University https://scholar.sun.ac.za

Due to the FCFS assumptions, the partial service rate solutions at a particular process can often be extended beyond the point where the arrival rates are known. If the arrival functions at node i are known over the interval $[0, t_i]$ and the node is overloaded at time t_i, then the corresponding service rates can be calculated for the interval $[0, g_i(t_i)]$. These partial solutions can then be used to extend the partial arrival rate solutions at processes where $r^p_{i,j} > 0$.

3.6 Summary and conclusion

The OPD has a complex queueing system which is difficult to analyse using traditional queueing theory models. The literature review in this chapter provides an overview of some existing research into queueing systems with similar properties, and discusses how these models are applicable to the OPD system. This discussion highlights certain important properties of the OPD queues, including

1. Non-stationary arrival distributions and staff levels;
2. Periods high traffic intensity and overloading ($\rho(t) \geq 1$);
3. A network of linked queueing systems;
4. Multiple types of patients with different routings, arrival and service distributions;
5. Queues with mixed priority/FCFS queueing disciplines.

The three models presented in this chapter describe the OPD queueing system in terms of a set of $n \times m$ sub-queue variables. Each of these variables, denoted by $q^p_i(t)$, describes the number of patients from profile $p \in \mathcal{P}$ in the queue at process $i \in \mathcal{I}$. The purpose of these variables is to describe both the length and composition of each of the OPD queues over the course of the day.

The first model (§ 3.3) is based on the Chapman-Kolmogorov equations for continuous time Markov processes. This model incorporates the sub-queue variables into a discrete, finite state space, which fully describes the range of possible sub-queue lengths in the OPD system. The discussion of this model focusses on calculating the coefficients needed to describe the rate at which the OPD system transitions between these different states.

The Markov model is theoretically interesting because it provides a probabilistic description of the OPD system. This requires a very large state space to incorporate all possible combinations of queue lengths, and a correspondingly large number of equations to describe the transitions between different states. Unfortunately, the resulting system of equations is too large to be practically implemented and solved.

The second and third models presented in this chapter (§ 3.4–§ 3.5) use a simpler representation of the OPD system, which is based on fluid approximation techniques for queueing networks. The fluid models give a deterministic description of the expected behaviour of the queueing network and approximate the discrete queue length variables with continuous variables.

Both fluid models make use of discontinuous differential equations to describe the flow of patients through each queue. These equations model the increase in queue length during periods of high traffic intensity, and calculate how staff divide their time between different types of patients in the queue. During periods of low traffic intensity, the fluid equations describe the flow of patients from one process to another, but they do not model the queue lengths.

The main difference between the two fluid approximation models is related to the OPD queueing disciplines. The first model focusses on the priority of patients within each queue, and assigns

Stellenbosch University https://scholar.sun.ac.za

a higher proportion of the staff to high priority profiles. The second model uses a strict FCFS queueing discipline which ensures that patients leave the queue in the order in which they arrived. Neither model is an exact representation of the OPD queues, which use a mixture of priority and FCFS queueing.

The fluid approximation models are both implemented in Mathematica (Wolfram Research, Inc.). The PF equations are solved numerically, while the FCFS equations are solved analytically. The computational efficiency of these approaches and the results of these models are discussed in Chapter 6.

Stellenbosch University https://scholar.sun.ac.za

Stellenbosch University https://scholar.sun.ac.za

CHAPTER 4

Simulation model

In this chapter, a simulation model for the OPD queueing network is presented. Instead of modelling the queue length at each process (as in the previous chapter), the simulation uses an agent-based modelling approach to follow the movements of individual patients within the OPD. Section 4.1 begins with a brief overview of the applications of agent-based modelling in Operations Research and healthcare, and the agent-based OPD model is presented in § 4.1.2. Section 4.2 describes how this model is implemented using a discrete event simulation (DES).

4.1 Agent-based modelling

Agent-based modelling first appeared in the 1990s, and it is still considered to be a relatively new technique. According to Siebers *et al.* (2010), the purpose of agent-based models is

> "[to] help us better understand real-world systems in which the representation or modelling of many individuals [agents] is important and for which the individuals have autonomous behaviours."

Although agent-based modelling is not widely used in Operations Research (Siebers *et al.*, 2010), certain authors have suggested that it is a very effective way to model a variety of OR problems. Borshchev & Filippov (2004) assert that a key advantage of these models is their distinctive *bottom-up* approach to large systems. Techniques that focus on the high-level behaviour of these systems often require very simplified representations of individual entities within the system, while agent-based models focus on the behaviour and interactions of these entities.

4.1.1 Literature

The agent-based approach is particularly relevant to healthcare facilities, where it is important to consider the behaviour and outcomes of individual patients within a much larger system (Stainsby *et al.*, 2009). Agent-based models can be used to capture the complex interactions between individuals in these facilities and to develop a better understanding of how these interactions affect the behaviour of staff and patients.

Much of the existing literature on agent-based models in healthcare systems focuses on disease transmission models for large healthcare facilities like hospitals, where infections can spread

Stellenbosch University https://scholar.sun.ac.za

rapidly through interactions between patients and staff in the facility. However, there are a few recent examples of agent-based models being used to assess and improve the efficiency of healthcare facilities.

Laskowski & Mukhi (2008) use an agent-based model to represent patients in multiple hospital emergency departments within the same city. The aim of this model is to investigate how the average queue length in each of these facilities is affected by various operational policies. The model is used to evaluate different staffing schedules and strategies for balancing the patient load at each facility by diverting incoming patients to nearby hospitals when a particular facility is overloaded.

A comprehensive agent-based model for a single hospital emergency department is described in Stainsby *et al.* (2009), Taboada *et al.* (2011), and Cabrera *et al.* (2011). This model includes multiple different types of agents such as doctors, nurses, technicians, patients, and relatives. It is used to investigate how patients' waiting times and outcomes are affected by different combinations of staff with varying levels of experience. This idea is extended in Mejıa-Quintero & Escudero-Marın (2015), which uses an agent-based approach to model the performance of doctors in an emergency department based on stress and fatigue levels.

4.1.2 OPD applications

The model introduced here uses an agent-based approach to model individual patients in the OPD and generate data regarding their progress through the system over the course of a day. Various performance measures such as waiting times and queue lengths can be calculated from this information.

The simulation procedure begins by generating a set of agents to represent patients that come to the OPD over the course of a single day. Each agent in the set is assigned to one of the m patient profiles, and the total number of agents for each profile is determined by selecting a random integer from the distribution $f_{\eta_p}(x)$.

Agents are characterised by a specific set of treatment data, which are randomly generated using the parameters in the OPD conceptual model. This includes **(a)** when the patient will arrive; **(b)** which processes the patient will go to; **(c)** the order in which they will go to these processes; and **(d)** their priority in these queues. Due to Assumption 3, Assumption 4 and Assumption 5 (§ 2.4), these characteristics do not change due to interactions with other patients and staff in the OPD, and so all of this information is generated at the beginning of the simulation (see Procedure 4.1).

Detailed information is collected during the simulation, such as the location of each patient throughout the day, how long they spend in each queue, and the time that they leave the OPD. Tracking each patient continuously is both unnecessary and computationally expensive, so the agent-based model is implemented as a discrete event simulation.

4.2 Discrete event simulation

Discrete event simulation is a fast, effective way to simulate systems that change over time in a series of discrete steps. These changes in the system (or any of its components) are referred to as *events*. The main assumption of DES models is that all events occur instantaneously, and there is no change in the system in the intervals between consecutive events.

Stellenbosch University https://scholar.sun.ac.za

An *event list* is used to keep track of all the events that are scheduled to take place in the simulation. The simulation evaluates events from this list in chronological order, updating the system accordingly after each event. In addition to changes to the state of the system, some events may result in new events being added to the event list or cancellation/postponement of existing events. After an event has been evaluated, it is removed from the event list, the event list is re-sorted, and the algorithm jumps forward to the occurrence of the next event.

4.2.1 OPD applications

In the agent-based OPD model, seven types of events will be considered:

1. A new patient arrives at the OPD (**A**)
2. A patient joins a queue (**J**)
3. A patient begins a treatment (**B**)
4. A patient completes a treatment (**C**)
5. A patient leaves the OPD (**L**)
6. A staff member starts a shift / returns from a break ($\mathbf{S}^+$)
7. A staff member ends a shift / goes off duty for a break ($\mathbf{S}^-$)

Most of these events will have immediate consequences, which are determined by the current state of the queues. For example, a new arrival (**A**) will either go to the queue at their first process (**J**) or begin treatment (**B**) if the queue is empty. Since the (**J**) or (**B**) event will occur directly after (**A**), it is evaluated immediately without being added to the event list.

The only events that have delayed consequences are the start of a new treatment (**B**), which results in the completion of this treatment (**C**) at some point in the future. Since the treatment times for each patient are generated at the beginning of the simulation (and independent of the state of the system), the time of the completion event is calculated and added to the event list as soon as the patient's treatment begins.

In practice, the only events that will appear on the event list are **A**, **C**, $\mathbf{S}^+$ and $\mathbf{S}^-$ events, as the remaining three events only occur as an immediate consequence of another event (see summary in Table 4.1). The possible outcomes of an event on the event list are as follows:

1. A new patient arrives at the OPD (**A**):

   The patient will go to their first process and either
   - join the queue (**J**), or
   - begin treatment immediately (**B**). The completion of this treatment will be added to the event list (**C**).

2. A patient completes a treatment (**C**):
   - If the patient has completed all necessary processes, they will leave (**L**).
   - Otherwise, the patient will go to their next process and either
     - join the queue (**J**), or

Stellenbosch University https://scholar.sun.ac.za

– begin treatment immediately (**B**). The completion of this treatment will be added to the event list (**C**).

The staff member who treated this patient will either

- wait for the next patient to arrive (if there is no queue), or
- go off duty if they are due to leave ($\mathbf{S}^-$), or
- begin treating the first patient in the queue (**B**). The completion of this treatment will be added to the event list (**C**).

3. A new staff member starts a shift / returns from a break ($\mathbf{S}^+$):

   The staff member will either

   - wait for the next patient to arrive (if there is no queue), or
   - begin treating the first patient in the queue at their process (**B**). The completion of this treatment will be added to the event list (**C**).

4. A staff member ends a shift / goes off duty for a break ($\mathbf{S}^-$):

   - If there is a staff member at this process who is not treating a patient, they will go off duty immediately.
   - If not, an event is added to the front of the queue for that process to indicate that the staff member should go off duty when they have finished with their current patient.

Trigger	Event	Possible consequences
event list	**A**	**J**, **B**
event list	$\mathbf{S}^+$	**B**
event list, **C**	$\mathbf{S}^-$	-
event list	**C**	(**B**, **L**, **J**) and (**B**, $\mathbf{S}^-$)
$\mathbf{S}^+$, **A**, **C**	**B**	add **C** to event list
A, **C**	**J**	-
C	**L**	-

TABLE 4.1: *A summary of the OPD simulation events with their triggers and consequences.*

4.2.2 Simulation algorithm

Pseudocode for the discrete event simulation is given in Procedures 4.1–4.7 and Algorithm 4.8. Table 4.2 provides a summary of these components, which are discussed in the remainder of this chapter.

	Pseudocode	Description
Data and initialisation:	Procedure 4.1	Generate patient sample
	Procedure 4.2	Define event types
	Procedure 4.3	Generate initial event list
Event evaluation:	Procedure 4.4	Completion event evaluation
	Procedure 4.5	Arrival event evaluation
	Procedure 4.6	Staff increase event evaluation
	Procedure 4.7	Staff decrease event evaluation
Simulation:	Algorithm 4.8	Run the discrete event simulation

TABLE 4.2: *A summary of the pseudocode for the OPD simulation model.*

Stellenbosch University https://scholar.sun.ac.za

The simulation algorithm begins with three initialisation procedures that use the OPD conceptual model parameters to generate data and events needed during the simulation. The first of these procedures (4.1) generates a sample of OPD patients, as discussed in § 4.1.2.

Procedure 4.2 defines four event classes to represent the **A**, **C**, $\mathbf{S}^+$ and $\mathbf{S}^-$ events. Instances of these event classes are used to store local variables t, p, i, and j, which describe the properties of particular events that take place in the simulation. These properties indicate the time of the event and the patients and processes that are affected by the event. Each event in the simulation is an instance of one of these event classes.

Procedure 4.3 combines the patient sample from Procedure 4.1 with the staff schedule parameters to generate an initial event list. This initial event list contains one type **A** event for every patient in the sample, and multiple type $\mathbf{S}^+$ and $\mathbf{S}^-$ events which correspond to increases and decreases in the number of staff on duty over the course of the day.

In addition to the event list, Procedure 4.3 also initialises n empty event queues which correspond to the n OPD processes. These event queues are used to store **C** and $\mathbf{S}^-$ events that cannot be placed on the event list because their time, t, depends on the evaluation of future events. In these queues, the **C** events represent the patients queueing at each process, while the $\mathbf{S}^-$ events represent staff members who are due to take a break or end their shift as soon as they finish treating their current patient.

Unlike the event list, which is sorted chronologically, the event queues are sorted by priority. The highest priority in these queues is 0, which is assigned to $\mathbf{S}^-$ events. The remaining **C** events are assigned a priority corresponding to the ϑ_i^p priority parameters.

When a staff member becomes available at a particular process, the highest priority event in the corresponding queue is selected. If this event is a staff decrease event, the number of available staff is reduced by one and the event is removed from the queue. If the highest priority event is a treatment event, the time of the event is updated to reflect when the treatment will be completed and the event is added to the event list. Once an event has been added to the event list, its time and other properties cannot be changed.

Algorithm 4.8 gives a simple framework for the discrete event simulation, which iteratively evaluates events on the event list using Procedures 4.4–4.7. These procedures use the local variables from each of the different event classes to represent the properties of the current event, and they modify the global simulation variables to reflect the consequences of the events. Evaluated events are removed from the event list, and the simulation algorithm terminates when the event list is empty.

Stellenbosch University https://scholar.sun.ac.za

Procedure 4.1: Generate patient sample

Inputs : $\mathcal{I}$, // processes
$\mathcal{P}$, // profiles
$f_{\eta_p}(x)$, // patient count distributions
$\alpha_p(x)$, // arrival time distributions
ϱ_i^p, // treatment needs probabilities
ϕ_i^p, // process routing order
ϑ_i^p, // priorities
$f_{\tau_i^p}(x)$, // treatment time distributions

Outputs: N, // number of patients
arrivalTime, // patient arrival times
treatmentNeeds, // a list of processes visited by each patient
treatmentTimes, // treatment times for patients at each process
priorities, // priorities for patients at each process

```
N = 0                                                          // Initialise η
for p ∈ P do
    η_p = random(f_{η_p}(x))                  // Generate number of patients for profile p
    for k ∈ {1, 2, ..., η_p} do
        N = N + 1
        arrivalTime[N] = random(α^p(x))
        for i ∈ I do
            if random(uniform(0,1)) ≤ ϱ_i^p then
                append i to treatmentNeeds[N]                  // Patient needs process i
                treatmentTimes[N,i] = random(f_{τ_i^p}(x))     // Generate treatment time
            end
        end
        sort treatmentNeeds[N] by [φ_1^p, φ_2^p, ..., φ_n^p]  // List processes in order
        priorities = [ϑ_1^p, ϑ_2^p, ..., ϑ_n^p]
    end
end
```

Procedure 4.2: Define event types for DES

Define A(t, p, i, j)**:**
- t = time of event
- p = patient
- i = patient's first process
- j = patient's remaining processes

end

Define C(t, p, i, j)**:**
- t = time of event
- p = patient
- i = current process
- j = patient's remaining processes

end

Define S$^+(t, i)$**:**
- t = time of event
- i = process

end

Define S$^-(t, i)$**:**
- t = time of event
- i = process

end

Stellenbosch University https://scholar.sun.ac.za

Procedure 4.3: Build initial event list

Inputs : $\mathcal{I}$, // processes
$\mathcal{P}$, // profiles
$\varsigma_i(t)$, // staff schedule functions
N, // number of patients
arrivalTime, // patient arrival times
treatmentNeeds, // a list of processes visited by each patient
treatmentTimes, // treatment times for patients at each process

Outputs: eventList, // initial event list

for $i \in \mathcal{I}$ **do**
  for $t \in \{30, 60, \ldots, 24 \times 60\}$ **do**
    if $\varsigma_i(t-30) < \varsigma_i(t)$ **then**
      $x = \varsigma_i(t) - \varsigma_i(t-30)$ // Additional staff on duty
      for $k \in \{1, \ldots, x\}$ **do**
        append $\mathbf{S}^+(t, i)$ to eventList
      end
    else if $\varsigma_i(t-30) > \varsigma_i(t)$ **then**
      $x = \varsigma_i(t-30) - \varsigma_i(t)$ // Staff go off duty
      for $k \in \{1, \ldots, x\}$ **do**
        append $\mathbf{S}^-(t, i)$ to eventList
      end
    end
  end
  $\mathsf{Q}_i = \emptyset$ // Initialise empty queue at process i
  $s_i = 0$ // Initialise number of staff available
end

// Make patient arrival events

for $p \in \{1, 2, \ldots, \mathsf{N}\}$ **do**
  $t =$ arrivalTime$[p]$
  $i =$ treatmentNeeds$[p, 1]$
  $j =$ treatmentNeeds$[p, 2 : end]$
  append $\mathbf{A}(t, p, i, j)$ to eventList
end

Stellenbosch University https://scholar.sun.ac.za

Procedure 4.4: Completion event evaluation

Global variables: eventList, treatmentTimes, priorities, Q_i, s_i
Local variables : t // time of event
p // patient
i // current process
j // patient's remaining processes

$s_i = s_i + 1$ // Increase available staff
if $j \neq \emptyset$ **then** // Patient does not leave the OPD
  $k = j[1]$ // Patient's next process
  $r =$ treatmentTimes$[p,k]$
  delete $j[1]$
  if $s_k \geq 1$ **then**
    $s_k = s_k - 1$ // Decrease available staff
    append $\mathbf{C}(t + r, p, k, j)$ to eventList
  else
    append $\mathbf{C}(r, p, k, j)$ to Q_k
  end
end
if $\mathsf{Q}_i \neq \emptyset$ **then** // Process event from Q_i
  sort Q_i by event priorities
  load $\mathsf{Q}_i[1]$
  if *current event type* $= \mathbf{S}^-(t^*, i)$ **then**
    $s_i = s_i - 1$ // Decrease available staff
  else if *current event type* $= \mathbf{C}(t^*, p^*, i, j^*)$ **then**
    $s_i = s_i - 1$ // Decrease available staff
    append $\mathbf{C}(t + t^*, p^*, i, j^*)$ to eventList
  end
  delete $\mathsf{Q}_i[1]$
end

Procedure 4.5: Arrival event evaluation

Global variables: eventList, treatmentTimes, priorities, Q_i, s_i
Local variables : t // time of event
p // patient
i // current process
j // patient's remaining processes

$r =$ treatmentTimes$[p,i]$
if $s_i \geq 1$ **then** // Patient is treated
  $s_i = s_i - 1$ // Decrease available staff
  append $\mathbf{C}(t + r, p, i, j)$ to eventList
else // Patient joins queue
  append $\mathbf{C}(r, p, i, j)$ to Q_i
end

Stellenbosch University https://scholar.sun.ac.za

Procedure 4.6: Staff increase event evaluation

Global variables: eventList, treatmentTimes, priorities, Q_i, s_i
Local variables : t // time of event
p // patient
i // current process
j // patient's remaining processes

```
1  s_i = s_i + 1                                              // Increase available staff
2  if Q_i ≠ ∅ then
3  |  load Q_i[1]                                            // Select first event from Q_i
4  |  if current event type = S^-(t*, i) then
5  |  |  s_i = s_i − 1                                       // Decrease available staff
6  |  else if current event type = C(t*, p*, i, j*) then
7  |  |  s_i = s_i − 1                                       // Decrease available staff
8  |  |  append C(t + t*, p*, i, j*) to event list
9  |  end
10 |  delete Q_i[1]
11 end
```

Procedure 4.7: Staff decrease event evaluation

Global variables: eventList, treatmentTimes, priorities, Q_i, s_i
Local variables : t // time of event
i // process

```
1  if s_i ≥ 1 then
2  |  s_i = s_i − 1                                           // Decrease available staff
3  else
4  |  append S^-(t, i) to Q_i
5  end
```

Algorithm 4.8: Discrete Event Simulation

```
1  call Procedure 4.1                                        // Generate patient sample
2  call Procedure 4.2                                        // Define event types
3  call Procedure 4.3                                        // Build initial event list
4  while eventList ≠ ∅ do
5  |  sort eventList
6  |  load eventList[1]
7  |  if current event type = C(t, p, i, j) then
8  |  |  call Procedure 4.4                                  // Completion event evaluation
9  |  else if current event type = A(t, p, i, j) then
10 |  |  call Procedure 4.5                                  // Arrival event evaluation
11 |  else if current event type = S^+(t, i) then
12 |  |  call Procedure 4.6                                  // Staff increase event evaluation
13 |  else if current event type = S^-(t, i) then
14 |  |  call Procedure 4.7                                  // Staff decrease event evaluation
15 |  end
16 end
```

Stellenbosch University https://scholar.sun.ac.za

Stellenbosch University https://scholar.sun.ac.za

CHAPTER 5

Data and parameters

This chapter presents two sets of parameters for the OPD conceptual model, which are referred to as *set-ups*. The first set-up applies to the OPD system at the beginning of 2015, while the second represents the system at the beginning of 2016. Both set-ups use the same patient profiles, but there are differences in the processes and treatment parameters which reflect the changes that were made in the OPD over the course of 2015.

Where possible, the model parameters have been estimated from data collected in the OPD. Section 5.1 gives an overview of two recent data collection projects and discusses some of the challenges associated with data collection. This data is used to identify the patient profiles in § 5.2 and parameters relating to the processes and staff in § 5.3.

Section 5.4 contains an overview of the treatment parameters in both the 2015 and 2016 set-ups. These parameters are based on a number of different sources, including data from the OPD, estimates provided by the OPD staff, and observations during visits to the OPD. Section 5.4 also considers the accuracy and reliability of these various sources.

5.1 Data collection

The OPD uses a paper-based system of handwritten patient records. It is very difficult to obtain accurate historical data from these records, since they are spread across many different sources. There are also several factors related to the OPD environment that make data collection very challenging. For example,

- There are no funds available to purchase equipment or employ additional staff to collect data.
- Overcrowding makes it difficult to keep clear, accurate records, since staff are more likely to make mistakes when they are rushing.
- Some patients do not have Identity Documents or records of their medical history.
- Many patients are illiterate and cannot fill in forms or identify medical documents.
- Staff who are not fluent in local languages must communicate with patients through interpreters, which makes it more difficult to obtain accurate information.

The OPD staff are aware that the lack of reliable data makes it difficult to identify and solve problems in the OPD, and two recent projects have attempted to address this issue. The first of

Stellenbosch University https://scholar.sun.ac.za

these was a pilot data collection project which took place at the end of 2014. This project was carried out in collaboration with Adam Bertscher, a Public Health student from the University of Cape Town. The main focus of this project was to identify sources of congestion in the OPD and collect qualitative and quantitative data regarding the effects of congestion on patient care.

According to Bertscher (2015), qualitative data gathered from interviews with OPD staff and patients highlighted several important issues that contributed to congestion in the OPD queueing process. Many patients were confused or frustrated by the OPD queues and reported that they had waited in the wrong queue at some stage during their OPD visit. This perspective was shared by the OPD nursing staff, who felt that staff were not providing clear instructions to patients about which queues to go to next.

Both nurses and doctors felt that the mixture of casualty and returning patients in a single facility was problematic. Doctors were concerned that they were unable to identify and treat very urgent patients promptly, because these patients would be sent to the same queue as everyone else. The nurses felt that the casualty patients took up a lot of space in the OPD and made it more difficult to process patients efficiently.

The quantitative data for this project was collected over a period of two days. Patients were given a form at the OPD entrance to track each step of their progress through the OPD. Staff members at each process were asked to write down the name of the process as well as the time that their interaction with the patient began and ended. These forms also required the patient's diagnosis or reason for visiting the OPD.

The quantitative data from these forms gives some insight into the type of patients in the OPD and how long they spent at different processes, but it also highlights several problems with this data collection strategy. Some patient forms went missing because the patients either lost the form in the OPD or took it with them when they left. Other forms were incomplete because some staff members did not record all the required information or completely forgot to ask for the patient's form.

Due to these problems, it is not possible to use this data to build patient profiles or to determine the routing of patients through the OPD. However, it does provide some information about the amount of time used to treat patients at different processes, which is helpful in determining the treatment time parameters.

The times recorded on these forms are not a true reflection of the rate at which patients are processed, because the forms do not include the time that elapses between consecutive patients. This changeover time between patients is especially relevant at the CLERKS and DOCTORS processes. The clerks spend a significant amount of time dealing with general enquiries from passing patients, while the doctors often need to clean up between consultations and spend a few minutes locating their next patient.

Bertscher (2015) identifies several ways in which the OPD queues could be improved. Since the DOCTORS queue was the main source of delays, doctors were encouraged to avoid spending time on tasks that could be delegated to other staff. The report also made several recommendations on how to streamline the administrative queues and simplify the system to avoid confusing patients.

At the end of 2015, a second set of data was collected as part of an audit of the casualty facilities at Zithulele. The purpose of this project was to gather data regarding the number, type and needs of casualty patients. The scale of this project was much larger, and data for more than four thousand casualty patients was collected over a period of three months.

Stellenbosch University https://scholar.sun.ac.za

The data collection procedure during the audit was also based on patient forms, but the information recorded on these forms was different to the 2014 project. There was a stronger emphasis on patient information, including age, gender, symptoms, diagnosis and outcome, and less attention was given to information about the different processes. Instead of recording the time at which a patient started and ended each process, only three times were noted: when the patient arrived, when they were triaged and when they were seen by a doctor.

The audit data is used to identify four different casualty profiles, which are discussed in § 5.2. Although the audit did not monitor non-casualty patients, some information about these patients is available in the OPD records. This data includes records of non-casualty arrivals spanning the first two months of the audit period.

5.2 Patient profiles

The following four casualty patient profiles were identified from the audit data:

- Green: casualty patients with minor injuries/illnesses.
- Yellow: casualty patients with moderate injuries/illnesses.
- Orange: casualty patients with serious injuries/illnesses.
- Red: casualty patients with very serious injuries/illnesses.

These four groups represent the triage system that the OPD uses to classify all casualty patients. The groups separate patients according to their urgency rather than their diagnosis, which means that each of these profiles contains patients with a wide variety of different treatment needs.

It is possible to create more homogeneous profiles by classifying patients according to the nature of their injury or illness (burns, fractures, bleeding, etc). However, these profiles would be less useful in the OPD model. The model focuses on waiting times, so the urgency of a patient's condition is more relevant than the type of injury/illness.

The triage profiles are also more informative because they are linked to specific waiting time targets, which limit the maximum amount of time that a patient should wait before being seen by a doctor. The OPD aims to treat green patients within 4 hours of their arrival, yellow patients within 1 hour, orange patients within 10 minutes and red patients within 2 minutes. These targets provide a way to assess how effectively the OPD is meeting the needs of patients in each of the casualty profiles.

The non-casualty OPD patients are divided into two groups:

- Sleepover: patients who have waited overnight at the OPD.
- Return: patients who come back regularly for long term/chronic conditions.

The sleepover profile can contain a mixture of returning and casualty patients from the previous day. They have often completed most of the required processes during the previous day and are waiting to collect test results or medication in the morning. These patients choose to wait at the OPD overnight, and are not admitted to the hospital. The target times for sleepover and returning patients are 30 minutes and 2 hours.

Each patient profile in the OPD model is associated with two sets of parameters which are discussed in the sections below. The first set of parameters gives the distribution of the number of patients per day in the OPD, while the second gives the distribution of the arrival times for patients in each profile.

Stellenbosch University https://scholar.sun.ac.za

5.2.1 Number of patients

According to the audit data, the OPD treats an average of 42 casualty patients and 36 returning patients between 7h00 and 17h00 on a normal weekday. Nearly half of the casualty patients are classified green (48.3%) and just over a third are yellow (35.4%). About 15% of casualty patients are orange and only 1.3% are red. During the audit, the hospital saw an average of 20.8 green patients per day, 14.4 yellow patients, 5.8 orange patients and 0.5 red patients. A plot of the distribution of daily patient counts recorded during the OPD audit is provided in Figure 5.1.

Based on this data, the number of patients per day is modelled as a triangular distribution with the probability density function

$$f(x) = \begin{cases} \dfrac{2(x-a)}{(c-a)(b-a)}, & a \leq x \leq c, \\ \dfrac{2(b-x)}{(b-a)(b-c)}, & c < x \leq b, \\ 0 & \text{otherwise}, \end{cases} \tag{5.1}$$

where a and b are the minimum and maximum values and c is the mode. The parameters for the different profiles are given in Table 5.1. Since the number of patients is a discrete random variable, the probability mass function for η_p is given by

$$f_{\eta_p}(x) = \mathrm{P}(\eta_p = x) = \int_{x-\frac{1}{2}}^{x+\frac{1}{2}} f(y)\mathrm{d}y, \quad x \in \{a, a+1, \ldots, b\}. \tag{5.2}$$

The distributions for the number of return, green, yellow and red patients are symmetric, while the orange patients' distribution is skewed to the right. The distribution of the number of red patients has been assigned a mode of 1, which over-estimates the number of days that the OPD treats at least one red patient. Although this is not representative of a typical week or month at the hospital, it provides a good indication of how well the hospital is able to deal with red patients on any given day.

Unfortunately, there is no data available on the number of sleepover patients at the OPD during the audit period. The staff at Zithulele estimate that there are usually between 5 and 15 of these patients, and so these numbers have been used as the lower and upper bounds in a symmetric triangular distribution for the number of sleepover patients.

5.2.2 Arrival times

The arrival times recorded during the 2015 audit are summarised as histograms in Figure 5.2. The data for returning patients was captured from the hospital records for the first two months of the audit period. Casualty patients who arrived outside of the OPD working hours are excluded, since these patients should be treated by the night shift staff. There is no arrival data available for sleepover patients, since they are already at the OPD before staff begin working in the morning.

Arrival times for returning patients and green patients follow a triangular distribution with a peak at about 9h00. This pattern is probably related to the availability of public transport to and from the hospital. Since these patients are not seriously ill, they tend to have planned their trip in advance and arranged to get to the hospital early in the day. The yellow and orange patients also follow a triangular pattern, although their arrivals are more evenly distributed and do not drop sharply at the end of the day.

Stellenbosch University https://scholar.sun.ac.za

The arrivals data collected during the audit may be somewhat misleading, as there is likely to be a delay between the patient's actual arrival time and their recorded arrival time during the peak arrival periods. Patients who arrive early start queueing before the OPD staff begin their shifts, so it takes a while to process this backlog at the beginning of the day. Although the recorded arrivals peak at about 9h00, most of these arrivals are likely to occur before 8h30, and the arrivals distributions for the OPD model are adjusted to reflect this. The parameters for the arrival distributions are given in the last three columns of Table 5.1.

Unfortunately, there is too little data available to draw any reasonable conclusions about the distribution of arrival times for red patients. Based on discussions with the OPD staff, their arrivals are assumed to follow a symmetric triangular distribution between 7h00 and 17h00.

		Number of patients			Arrival times		
Profile	Target	Min	Max	Mode	Start	End	Mode
Return	02h00	15	55	35	7h15	13h00	8h24
Green	04h00	5	35	20	7h15	16h30	8h38
Yellow	01h00	5	25	15	7h15	17h00	9h12
Orange	00h10	0	13	5	7h00	17h00	9h30
Red	00h02	0	2	1	7h00	17h00	12h00
Sleepover	00h30	5	15	10	7h20	7h30	7h25

TABLE 5.1: *A summary of the parameters for the OPD patient profiles.*

Stellenbosch University https://scholar.sun.ac.za

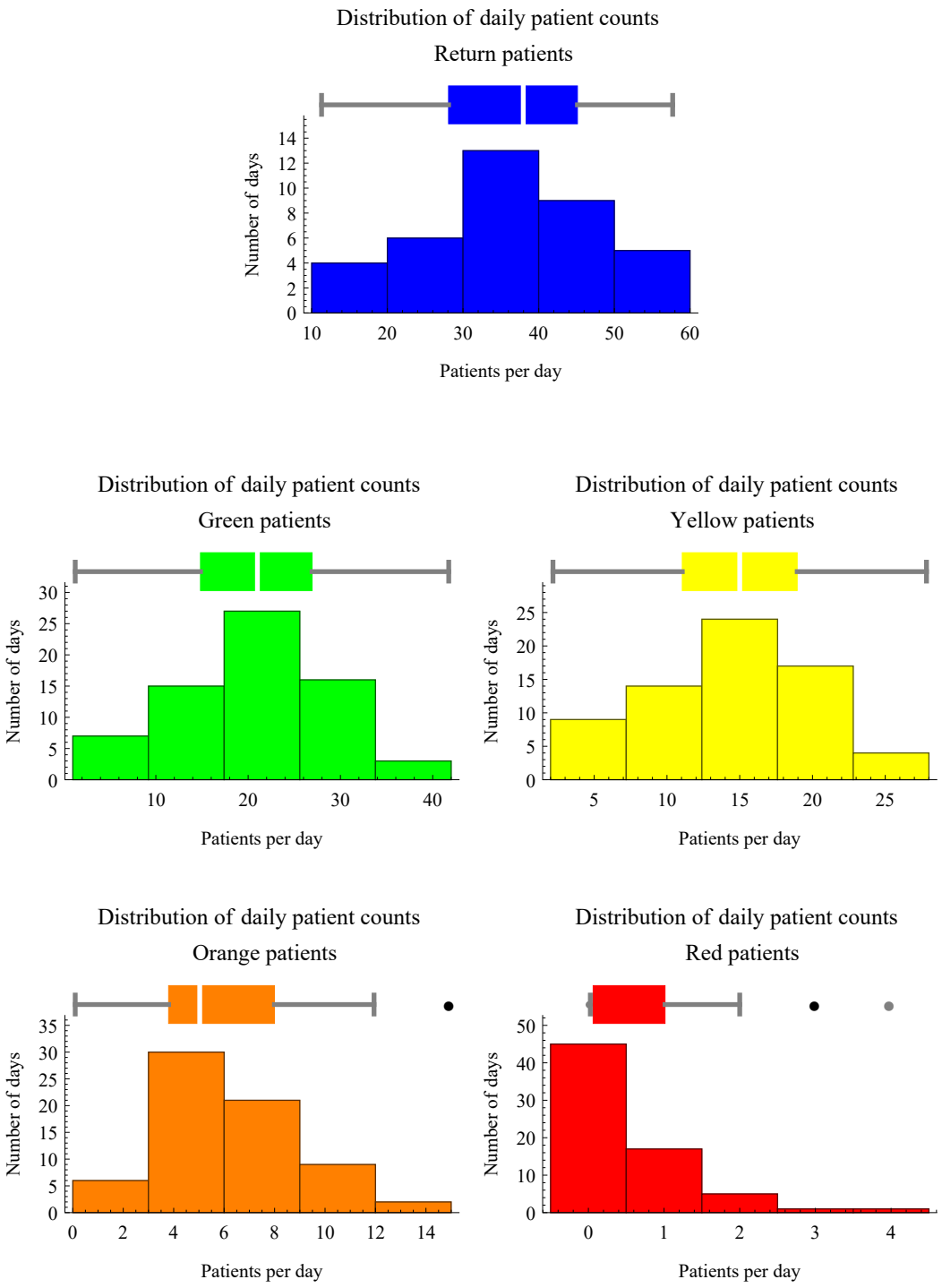


FIGURE 5.1: *Histograms of the number of patients per day recorded during the 2015 OPD audit. The box-and-whisker charts above each plot indicate the minimum, first quartile, mean, third quartile, and maximum of the observations.*

Stellenbosch University https://scholar.sun.ac.za

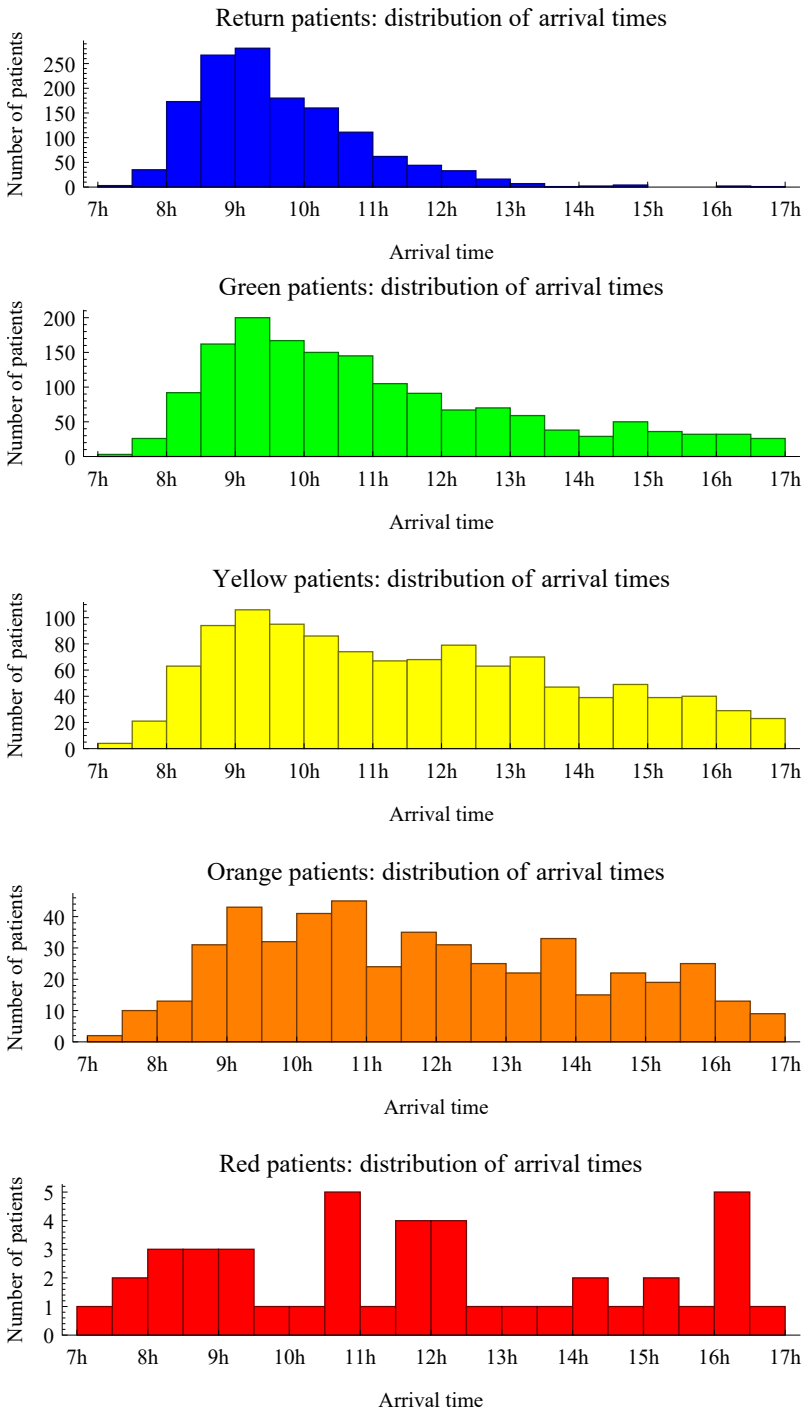


FIGURE 5.2: *The distribution of patient arrival times recorded during the 2015 OPD audit.*

Stellenbosch University https://scholar.sun.ac.za

5.3 Processes

Over the course of 2015, certain changes were made to the OPD processes to improve the way that different patient profiles are treated. Two different set-ups are used to represent the OPD before and after these changes. In § 5.3.1 the first set-up is explained, along with an overview of the basic structure of the OPD processes. The changes that have been made to this system for the second OPD set-up are described in § 5.3.2.

5.3.1 OPD processes in 2015

The 2015 OPD set-up represents a one-size-fits-all system where there is very little attempt to differentiate between patient profiles. All of the queues contain a mixture of casualty and returning patients who are treated on a FCFS basis.

There are six processes represented in the 2015 OPD set-up:

Process 1: CLERKS

Process 2: VITALS

Process 3: DOCTORS

Process 4: BLOOD TESTS

Process 5: X-RAYS

Process 6: PHARMACY

The normal procedure for an OPD patient is to begin with the administrative processes. On arrival, patients go to the CLERKS who record their details and stamp their patient books. Patients then proceed to a second queue (VITALS) to have their vital signs checked and be triaged (if necessary). These two steps are compulsory unless the patient is in a critical condition.

After these two processes, patients may go to different queues depending on their treatment needs: some go directly to the DOCTORS queue, while others must first queue for diagnostic tests (BLOOD TESTS and X-RAYS). Patients might also be admitted to the main hospital for further treatment or sent for specialist treatments such as physiotherapy or dental care. These specialist queues are not included in the model because they are not considered to be part of the main OPD system. Once a patient completes all the necessary tests and treatments they can collect medication at the PHARMACY before leaving the OPD.

One of the major problems with the 2015 OPD system is that many casualty patients are not treated within the appropriate time period. The main cause of this problem is the FCFS DOCTORS queue, which requires casualty patients to stand in the same queue as returning patients. This contributes to longer delays for casualty patients because they tend to arrive slightly later, once a backlog has already developed at this process.

Another problem with the 2015 OPD system is that it does not make allowances for the needs of urgent casualty patients who require immediate medical attention. In theory, red patients are allowed to bypass other patients in the DOCTORS queue, but this can only happen if they are correctly identified when they arrive in the OPD. If these patients do not have very obvious symptoms, they might only be identified once they have worked their way through the CLERKS, VITALS and DOCTORS queues.

During 2015, several changes were made to address these problems in the OPD. The aim of these changes was to streamline the treatment of casualty patients by separating them from the

Process	7h30	8h	8h30	9h	9h30	10h	10h30	11h	11h30	12h	12h30	13h	13h30	14h	14h30	15h	15h30	16h	16h30	17h	17h30	18h
CLERKS	3	3	3	3	3	2	2	2	3	3	2	2	2	2	2	2	3	3	3	3	1	1
VITALS	2	2	2	2	2	1	1	2	2	2	1	1	1	1	2	2	2	2	2	2	1	1
DOCTORS	0	0	3	3	3	2	2	2	3	3	2	2	2	2	2	3	3	4	4	4	1	1
BLOOD TESTS	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
X-RAYS	0	2	2	2	1	1	2	2	2	1	1	1	1	2	2	2	2	2	2	1	0	0
PHARMACY	0	0	0	2	2	2	2	1	1	2	2	2	0	0	2	2	2	2	2	2	2	0

TABLE 5.2: *The 2015 OPD staff schedule.*

Process	7h30	8h	8h30	9h	9h30	10h	10h30	11h	11h30	12h	12h30	13h	13h30	14h	14h30	15h	15h30	16h	16h30	17h	17h30	18h
CLERKS	3	3	3	3	3	2	2	2	3	3	2	2	2	2	2	2	3	3	3	3	1	1
VITALS	2	2	2	2	2	1	1	2	2	2	1	1	1	1	2	2	2	2	2	2	1	1
TRIAGE	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
CASUALTY DOCTORS	0	0	3	2	2	2	2	2	2	1	1	1	1	1	1	2	2	2	2	2	2	2
BLOOD TESTS	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
X-RAYS	0	2	2	2	1	1	2	2	2	1	1	1	1	2	2	2	2	2	2	1	0	0
RETURN DOCTORS	0	0	2	2	2	2	2	2	2	1	1	1	1	2	2	2	2	2	2	2	0	0
PHARMACY	0	0	2	2	2	2	1	1	2	2	2	0	0	2	2	2	2	2	2	2	0	0

TABLE 5.3: *The 2016 OPD staff schedule.*

Stellenbosch University https://scholar.sun.ac.za

Stellenbosch University https://scholar.sun.ac.za

returning patients and allowing doctors to prioritise the treatment of urgent cases. The next section describes the steps that were taken to facilitate these changes.

5.3.2 OPD processes in 2016

In the 2016 OPD set-up, two new processes are added the system to allow casualty patients faster access to medical treatments. Casualty patients no longer join the same VITALS or DOCTORS queue as the returning patients. Instead, they go directly from the CLERKS to a separate TRIAGE process where their vital signs are recorded. The staff at this process also classify patients into one of the four casualty groups based on a preliminary assessment of their condition, which helps to ensure that urgent patients are correctly identified.

Once casualty patients have been triaged they are seen by separate CASUALTY DOCTORS. This queue is shorter than the combined DOCTORS queue in the 2015 model, since it does not involve any returning patients. The queue for the CASUALTY DOCTORS is processed in order of urgency rather than on a FCFS basis, so patients in the red and orange groups experience fewer delays.

In this system, there are a total of eight processes:

Process 1: CLERKS

Process 2: VITALS (return only)

Process 3: TRIAGE (casualty only)

Process 4: CASUALTY DOCTORS (casualty only)

Process 5: BLOOD TESTS

Process 6: X-RAYS

Process 7: RETURN DOCTORS (return only)

Process 8: PHARMACY

The staff schedules for the 2015 and 2016 OPD set-ups are given in Table 5.3. There are no differences in the staff schedules for the CLERKS, VITALS, BLOOD TESTS, X-RAYS, and PHARMACY processes. The doctors' schedule changes to incorporate the separate casualty queue: the number of doctors treating returning patients decreases and additional doctors are available in the morning to treat casualty patients.

5.4 Treatment parameters

The purpose of the treatment parameters in the OPD set-ups is to describe the interactions between different patient profiles and the OPD processes. They indicate which processes each patient will go to, the order in which they visit each process, their treatment times at each process and their priority in the different process queues. The notation for these parameters is summarised in Table 5.4. In the rest of this section, each of these concepts is discussed in detail and an explanation of how the corresponding parameters were obtained is provided.

Tables 5.5 and 5.6 give a summary of the treatment parameters in the 2015 and 2016 OPD set-ups. The parameters for these set-ups reflect the same basic needs and characteristics of each patient profile, and the differences between the two sets of parameters reflect the changes implemented in the 2016 OPD set-up. The most significant difference between the 2015 and 2016 treatment parameters is that the casualty parameters for the VITALS and DOCTORS processes in

Stellenbosch University https://scholar.sun.ac.za

the 2015 set-up are moved to new TRIAGE and CASUALTY DOCTORS processes in the 2016 set-up. The 2016 queues also have different priority parameters which allow more urgent patients to be seen first.

Parameter	Explanation
ϱ_i^p	The proportion of patients in profile p who need process i.
mean(τ_i^p)	The average treatment time (in minutes) for patients in profile p who need process i.
range(τ_i^p)	The distance of the upper and lower bounds from the mean treatment times (in minutes).
ϑ_i^p	The order in which different patient profiles are treated.

TABLE 5.4: *A summary of the treatment parameters for patients in the OPD model.*

Profile: Sleepover

Process	ϱ_i^1	mean(τ_i^1)	range(τ_i^1)	ϑ_i^1
CLERKS	1	5m00	3m00	1
VITALS	1	5m00	3m00	1
BLOOD TESTS	0.2	10m00	5m00	3
X-RAYS	0.2	15m00	5m00	3
DOCTORS	1	10m00	5m00	2
PHARMACY	0.6	7m30	2m30	3

Profile: Return

Process	ϱ_i^2	mean(τ_i^2)	range(τ_i^2)	ϑ_i^2
CLERKS	1	5m00	3m00	1
VITALS	1	5m00	3m00	1
BLOOD TESTS	0.4	10m00	5m00	3
X-RAYS	0.4	15m00	5m00	3
DOCTORS	0.1	10m30	5m00	2
PHARMACY	1	6m00	4m00	3

Profile: Green

Process	ϱ_i^3	mean(τ_i^3)	range(τ_i^3)	ϑ_i^3
CLERKS	1	8m30	6m30	1
VITALS	1	7m00	4m00	1
DOCTORS	1	10m00	5m00	2
BLOOD TESTS	0.2	10m00	5m00	3
X-RAYS	0.3	15m00	5m00	3
PHARMACY	0.95	7m30	2m30	3

Profile: Yellow

Process	ϱ_i^4	mean(τ_i^4)	range(τ_i^4)	ϑ_i^4
CLERKS	1	10m00	5m00	1
VITALS	1	7m00	4m00	1
DOCTORS	1	14m30	6m00	2
BLOOD TESTS	0.4	10m00	5m00	3
X-RAYS	0.4	15m00	5m00	3
PHARMACY	0.9	7m30	2m30	3

Profile: Orange

Process	ϱ_i^5	mean(τ_i^5)	range(τ_i^5)	ϑ_i^5
CLERKS	0.8	10m00	5m00	1
VITALS	1	9m00	5m00	1
DOCTORS	1	32m00	15m00	2
BLOOD TESTS	0.8	10m00	5m00	2
X-RAYS	0.8	15m00	5m00	2
PHARMACY	0.5	10m00	5m00	2

Profile: Red

Process	ϱ_i^6	mean(τ_i^6)	range(τ_i^6)	ϑ_i^6
CLERKS	0.1	10m00	5m00	1
VITALS	1	10m30	5m00	1
DOCTORS	1	39m00	20m00	1
BLOOD TESTS	1	10m00	5m00	1
X-RAYS	0.8	17m30	7m30	1
PHARMACY	0.2	10m00	5m00	1

TABLE 5.5: *The treatment parameters for the 2015 OPD set-up.*

Stellenbosch University https://scholar.sun.ac.za

Profile: Sleepover

Process	ϱ_i^1	mean(τ_i^1)	range(τ_i^1)	ϑ_i^1
CLERKS	1	5m00	3m00	1
VITALS	1	5m00	3m00	1
TRIAGE	0	-	-	-
DOCTORS (C)	0	-	-	-
BLOOD TESTS	0.2	10m00	5m00	3
X-RAYS	0.4	15m00	5m00	3
DOCTORS (R)	1	10m00	5m00	1
PHARMACY	0.6	7m30	2m30	2

Profile: Return

Process	ϱ_i^2	mean(τ_i^2)	range(τ_i^2)	ϑ_i^2
CLERKS	1	5m00	3m00	1
VITALS	1	5m00	3m00	2
TRIAGE	0	-	-	-
DOCTORS (C)	0	-	-	-
BLOOD TESTS	0.4	10m00	5m00	4
X-RAYS	0.4	15m00	5m00	4
DOCTORS (R)	1	10m30	5m00	2
PHARMACY	1	6m00	4m00	2

Profile: Green

Process	ϱ_i^3	mean(τ_i^3)	range(τ_i^3)	ϑ_i^3
CLERKS	1	8m30	6m30	1
VITALS	0	-	-	-
TRIAGE	1	7m00	4m00	1
DOCTORS (C)	1	10m00	5m00	4
BLOOD TESTS	0.2	10m00	5m00	5
X-RAYS	0.3	15m00	5m00	5
DOCTORS (R)	0	-	-	-
PHARMACY	0.95	7m30	2m30	2

Profile: Yellow

Process	ϱ_i^4	mean(τ_i^4)	range(τ_i^4)	ϑ_i^4
CLERKS	1	10m00	5m00	1
VITALS	0	-	-	-
TRIAGE	1	7m00	4m00	1
DOCTORS (C)	1	14m30	6m00	3
BLOOD TESTS	0.4	10m00	5m00	5
X-RAYS	0.4	15m00	5m00	5
DOCTORS (R)	0	-	-	-
PHARMACY	0.9	7m30	2m30	2

Profile: Orange

Process	ϱ_i^5	mean(τ_i^5)	range(τ_i^5)	ϑ_i^5
CLERKS	0.8	10m00	5m00	1
VITALS	0	-	-	-
TRIAGE	1	9m00	5m00	1
DOCTORS (C)	1	32m00	15m00	2
BLOOD TESTS	0.8	10m00	5m00	2
X-RAYS	0.8	15m00	5m00	2
DOCTORS (R)	0	-	-	-
PHARMACY	0.5	10m00	5m00	2

Profile: Red

Process	ϱ_i^6	mean(τ_i^6)	range(τ_i^6)	ϑ_i^6
CLERKS	0.1	10m00	5m00	1
VITALS	0	-	-	-
TRIAGE	1	10m30	5m00	1
DOCTORS (C)	1	39m00	20m00	1
BLOOD TESTS	1	10m00	5m00	1
X-RAYS	0.8	17m30	7m30	1
DOCTORS (R)	0	-	-	-
PHARMACY	0.2	10m00	5m00	1

TABLE 5.6: *The treatment parameters for the 2016 OPD set-up.*

5.4.1 Treatment needs

The treatment needs parameters indicate which OPD processes a patient will go to, and which processes they will skip. Since treatment needs vary between different patients, this information is expressed in terms of a probability, ϱ_i^p, which represents the fraction of patients from profile p that need process i. These parameters are given in the first columns of Table 5.5 and Table 5.6.

Some of these parameters are easily identified because they are based on clear rules and procedures. For example, the CLERKS, VITALS and DOCTORS processes in the 2015 set-up are compulsory for most profiles. In these cases, a value of $\varrho_i^p = 1$ indicates that all patients in profile p must go to process i.

In the 2016 set-up, some patient profiles will never go to certain processes because there are separate queues for returning and casualty patients. The corresponding parameters are assigned a value of $\varrho_i^p = 0$.

Stellenbosch University https://scholar.sun.ac.za

It is more difficult to assign a value to ϱ_i^p when there is no clear-cut rule about whether a particular patient will need a certain process. Estimates for these parameters are based on the following factors:

- Types of injuries/illnesses
  The types of injuries and illnesses that are common in each profile influence the percentage of these patients that require BLOOD TESTS, X-RAYS and medication from the PHARMACY. Patients who do not need these processes will skip these queues.
- Severity of a patient's condition
  Red and orange patients sometimes skip compulsory processes like the CLERKS due to the seriousness of their condition.

The treatment needs parameter estimates used in the 2015 and 2016 OPD set-ups were obtained from the OPD staff. In the audit data, the actual number of patients with recorded BLOOD TESTS and X-RAYS is lower than these estimates. These discrepancies could be due to missing data or inconsistencies in the way that different staff members recorded these details.

It is also possible that the staff's estimates for the treatment needs parameters are based on the way that patients should be treated in ideal circumstances, which does not necessarily correspond with the way things are done in reality. For example, some patients who are supposed to have BLOOD TESTS or X-RAYS may skip these tests due to time constraints. Since it is difficult to determine the reliability and accuracy of the audit records, the OPD staff's higher parameter estimates are used in the 2015 and 2016 set-ups.

5.4.2 Routing

In Table 5.5, the OPD processes are listed in the order that they are visited by patients from each profile. The routing order is unchanged in Table 5.6 — the new casualty processes simply replace the original combined queues for certain profiles. In the 2015 set-up, patients begin with the CLERKS and VITALS queues and end with the PHARMACY (if necessary), but there are different procedures for casualty and returning patients when it comes to BLOOD TESTS and X-RAYS.

After the VITALS queue, casualty patients go directly to the DOCTORS queue and then proceed to the BLOOD TESTS or X-RAYS queues. Patients cannot have any diagnostic tests before they have seen a doctor, since the doctor needs to determine which tests are necessary. By contrast, returning patients who need BLOOD TESTS or X-RAYS can have these tests before they go to the DOCTORS. Based on the information in their patient book, staff can determine which tests they need when their vital signs are checked. These patients are generally familiar with the normal procedures for their check-ups and aware of which queues they need to stand in.

5.4.3 Treatment times

The treatment time parameters were particularly difficult to determine because there is no reliable data regarding the amount of time needed to treat patients at each OPD process. In the 2015 and 2016 OPD set-ups, the treatment times for particular profiles are described by two sets of parameters: a mean parameter, which gives the average treatment time, and a range parameter, which gives the lower and upper bounds for treatment times around the mean. The treatment times are modelled as symmetric triangular distributions between these upper and lower bounds.

Stellenbosch University https://scholar.sun.ac.za

The lack of data makes it difficult to identify the actual distributions of the service times. The motivation for using triangular service time distributions, rather than exponential, is that it is easier to discuss treatment times with the OPD staff in terms of a range, rather than a single mean value. In cases where the average treatment times are over- or under-estimated, the lower and upper bounds of the triangular distribution help to limit the treatment times for individual patients to a range of realistic values.

Although the OPD staff are a valuable source of information, treatment time parameter estimates based on their observations are not necessarily reliable. There are a number of factors which could influence the accuracy of these estimates, for example:

- Staff may fail to account for time spent changing between patients, finding patient records, and locating equipment needed to treat certain patients.
- Staff are more likely to remember treatments that were particularly difficult or time-consuming.
- Staff's estimates of treatment times may be skewed by the amount of time that they think they should be spending with each patient.

Due to these concerns, certain treatment time parameter estimates are adjusted based on information gathered during the 2015 audit. Although the audit data does not indicate the duration of a patient's treatment, it does provide the time that each patient was triaged and the time that they were seen by a doctor. Combined with the staff schedules, this data can be used to estimate how quickly different casualty patients are treated at these two processes.

For each weekday during the audit period, the audit data is used to calculate how many patients were processed during one-hour intervals from 9h00 to 17h00. Summaries of this data were used to identify busy periods with a higher than average ratio of patients to staff. The busy period for the TRIAGE queue runs from 9h00 until 13h00, and the busy period for the DOCTORS was from 10h00 to 16h00. During these periods, it is assumed that there was a constant stream of patients through these processes and no idle time between consecutive patients.

The treatment data for intervals during the busy periods were combined with information from the staff schedules to create a new dataset with five variables. Each instance in the dataset represents one of these intervals on a particular day during the audit period. The first four variables, x_1, x_2, x_3, and x_4 are the number of green, yellow, orange and red patients treated during that interval, and the fifth variable, s, gives the number of staff that were on duty during that interval. In cases where the number of staff changed during the interval, s is the average number of staff. The average treatment times for each profile are taken from the coefficients $\beta_1, \beta_2, \beta_3$ and β_4 in the regression equation

$$60s = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon. \tag{5.3}$$

Both regression models produced good results, with adjusted $R^2 = 0.76$ for the TRIAGE model and adjusted $R^2 = 0.86$ for the DOCTORS model. The coefficient estimates and other statistics are provided in Tables 5.7 and 5.8, and the residual plots are shown in Figures 5.3 and 5.4.

The residual plots indicate that the assumption of a linear relationship between the number of staff on duty and the number of patients treated does not necessarily hold during intervals that are unusually busy or unusually quiet. This is expected, since staff are likely to adjust the pace of their work depending on the length of the queues. The downward slope in the residual plots is most obvious for green and yellow patients because these are the biggest casualty profiles.

This problem can be fixed by transforming the variables x_1 and x_2, but this would change

Stellenbosch University https://scholar.sun.ac.za

	TRIAGE			DOCTORS		
Profile	Coefficient	SE	P-value	Coefficient	SE	P-value
Green	6.9	0.3	$< 10^{-6}$	9.4	0.8	$< 10^{-6}$
Yellow	7.2	0.5	$< 10^{-6}$	14.3	1.2	$< 10^{-6}$
Orange	8.8	1	$< 10^{-6}$	31.6	2.3	$< 10^{-6}$
Red	10.4	4.1	0.012	38.6	11	0.0005

TABLE 5.7: *The regression coefficients for the mean treatment times at the* TRIAGE *and* DOCTORS *processes.*

	TRIAGE	DOCTORS
Multiple R	0.93	0.87
R^2	0.87	0.76
Adjusted R^2	0.86	0.76

TABLE 5.8: *The regression statistics for the mean treatment times at the* TRIAGE *and* DOCTORS *processes.*

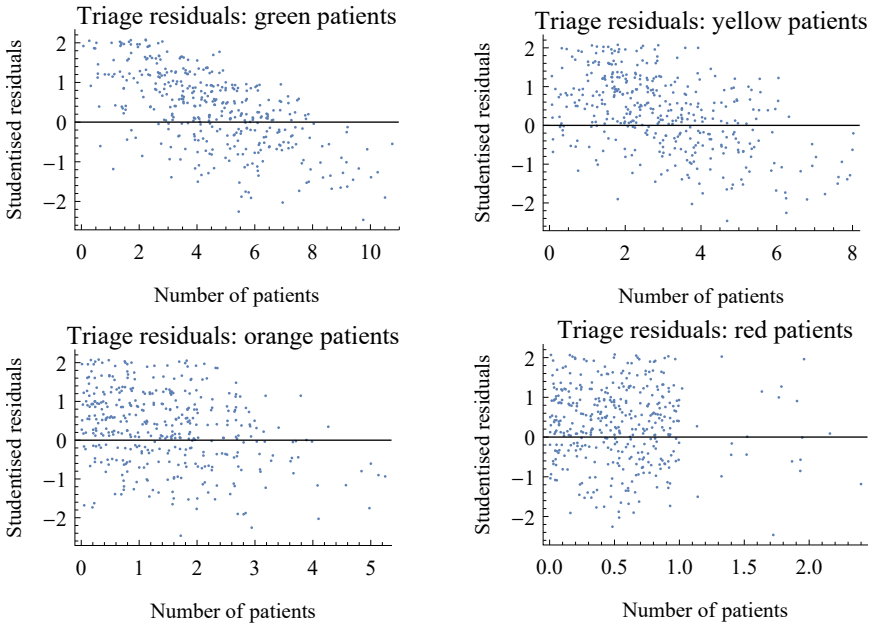


FIGURE 5.3: *Residual plots for the* TRIAGE *treatment times*[1].

the interpretation of the coefficients β_1 and β_2. For the purpose of parameter estimates, the assumption that treatment times are independent of the overall queue length is maintained.

The treatment time parameters obtained from the regression models are similar to the estimates given by the OPD staff. Since the audit data only applies to the TRIAGE and DOCTORS processes, there is still some uncertainty regarding the accuracy of the other treatment time parameters. The consequences of this uncertainty will be investigated through a sensitivity analysis in Chapter 6.

[1]Discrete x-axis values are scattered to ensure that all points are visible. The co-ordinates of plotted points $\mathbf{p}' = (x', y)$ correspond to data points $\mathbf{p} = (x = \lfloor x' \rfloor, y)$.

Stellenbosch University https://scholar.sun.ac.za

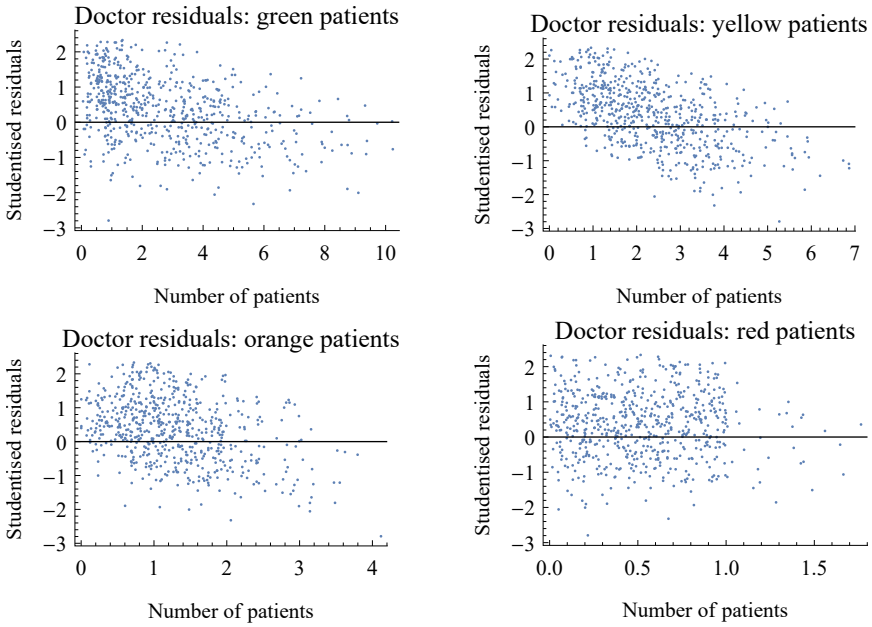


FIGURE 5.4: *Residual plots for the* DOCTORS *treatment times*[1].

5.4.4 Priority

The last column of the treatment parameter tables gives the queue priority (ϑ_i^p) for different patient profiles at each of the OPD processes. A priority of 1 indicates that a specific profile is treated first, and tied priorities for multiple profiles mean that these patients are treated on a FCFS basis.

Very little attention is given to prioritising different types of patients in the 2015 set-up. Red patients are allowed to move directly to the front of certain queues, since they often require immediate attention, but the remaining casualty and returning patients are seen in order of arrival.

In the 2016 OPD set-up there is a much greater emphasis on queue priorities, particularly for casualty patients. The 2016 queues for CASUALTY DOCTORS, BLOOD TESTS, and X-RAYS are restructured to treat profiles in order of their waiting time targets. Red patients are treated first, followed by orange, yellow, return and green patients. Sleepover patients are treated first in the queues for the CLERKS, VITALS, and RETURN DOCTORS.

5.5 Summary and conclusion

Data collection in the OPD is a time-consuming task, and it is very difficult to obtain accurate, detailed information about the OPD patients. Although the OPD management and staff are taking steps to address this problem, their efforts are limited by resource constraints, hospital infrastructure, and the congestion in the OPD facility.

Based on data collected during the 2015 OPD audit, this chapter divides the OPD patients into six different profiles. Casualty patients are classified as either green, yellow, orange or red according to their level of urgency, while return patients are grouped in a single profile. A

Stellenbosch University https://scholar.sun.ac.za

separate profile is used to represent sleepover patients who wait at the OPD overnight.

There are six OPD processes considered in this chapter: CLERKS, VITALS/TRIAGE, DOCTORS, BLOOD TESTS, X-RAYS, and PHARMACY. Two different OPD set-ups are used to illustrate the changes that were made to some of these processes during 2015. The 2015 OPD set-up represents the initial OPD system, where each of the six processes is associated with a single queue. In the 2016 OPD set-up, the VITALS and DOCTORS queues are split into separate processes for returning and casualty patients, resulting in a total of eight processes.

Due to these changes, the treatment parameters for the 2015 and 2016 set-ups are slightly different. The treatment needs parameters (ϱ_i^p) in the 2016 set-up are adjusted to ensure that the new VITALS/TRIAGE and CASUALTY DOCTORS/RETURN DOCTORS queues are restricted to either casualty or returning patients, and the routing of different patient profiles is updated accordingly. There are also changes to the priority disciplines at certain processes in the 2016 set-up to allow urgent patients to be seen more quickly.

Stellenbosch University https://scholar.sun.ac.za

Stellenbosch University https://scholar.sun.ac.za

CHAPTER 6

Results

This chapter presents the results of the queueing theory and simulation models for the OPD system, focusing on three aspects of these results:

1. The length of each queue over the course of the day.
2. The waiting times for different patient profiles.
3. Differences between the results for the 2015 and 2016 set-ups.

A discussion of the simulation results is provided in § 6.1, and the results of the fluid approximation models are presented in § 6.2.

6.1 Simulation model results

The purpose of this section is to demonstrate how the OPD simulation model can be used to develop a better understanding of congestion in the OPD system. Detailed data generated by the simulation model are used to analyse and compare the efficiency of the 2015 and 2016 OPD set-ups in § 6.1.1–§ 6.1.3, followed by a discussion of the accuracy and limitations of the simulation results in § 6.1.4–§ 6.1.5.

6.1.1 Number of simulations

The number of simulation runs required to analyse a particular OPD set-up depends on three factors:

1. The type of information being measured.
2. The variance in the simulation results.
3. The level of accuracy required.

For the results presented in § 6.1.2–§ 6.1.3, the number of simulation runs was calculated based on the average total waiting times for each of the patient profiles, as well as the average waiting times for each profile in each of the different OPD queues. These calculations provide a set of lower bounds N_i^p, which indicate the minimum number of simulation runs needed to estimate the average waiting time for profile p patients at process i.

Stellenbosch University https://scholar.sun.ac.za

The formula used to determine these lower bounds is

$$N_i^p > \frac{1}{\eta_p \varrho_i^p} \left(\frac{Z_{(1-\frac{\alpha}{2})} \hat{S}}{\epsilon} \right)^2, \qquad \text{with } i \in \mathcal{I} \text{ and } p \in \mathcal{P}, \tag{6.1}$$

where ϵ is the acceptable margin of error, $(1-\alpha)$ is the required confidence level and $\hat{S}$ is the estimated standard deviation of the waiting times for profile p patients at process i in the simulation results. This formula is based on the confidence interval method (Banks *et al.*, 2004; Robinson, 2004), and the terms $\eta_p \varrho_i^p$ are included to account for the number of patients from each profile p that are expected to go to each process i during each simulation run. A similar formula is applied to the average total waiting times for each of the different patient profiles to get a second set of lower bounds,

$$N_\Sigma^p > \frac{1}{\eta_p} \left(\frac{Z_{(1-\frac{\alpha}{2})} \hat{S}}{\epsilon} \right)^2, \qquad \text{with } p \in \mathcal{P}. \tag{6.2}$$

Table 6.1 contains the minimum sample sizes N_i^p and N_Σ^p that were calculated for both the 2015 and 2016 set-ups using the parameters $\epsilon = 1$ minute and $(1-\alpha) = 0.95$. In many cases, the number of simulation runs required is different in the 2015 and 2016 set-ups. This indicates that the amount of variability in these waiting times increases or decreases in response to the changes that are implemented in the 2016 set-up. Based on these results, a total of 2000 simulation runs were performed for both the 2015 and 2016 OPD set-ups.

The results in Table 6.1 indicate that relatively few simulations are needed to estimate the waiting times for return and sleepover patients. This is related to the fact that return patients are the biggest profile, while sleepover patients are the most predictable profile due to their early arrival times.

	Return		Green		Yellow		Orange		Red		Sleepover	
	2015	2016	2015	2016	2015	2016	2015	2016	2015	2016	2015	2016
CLERKS	1	1	2	2	2	2	8	8	301	301	3	3
VITALS	6	1	10	15	13	18	41	58	354	27	2	2
DOCTORS	63	40	146	340	217	107	649	64	89	160	7	12
BLOOD TESTS	23	29	58	791	37	473	43	6	18	17	29	28
X-RAYS	26	20	62	399	62	347	78	9	38	32	52	50
PHARMACY	11	17	22	29	33	49	182	225	1967	1739	1	3
All processes	70	142	455	243	350	356	239	1012	294	376	24	19

TABLE 6.1: *The number of simulations required to estimate mean waiting times with $\alpha = 0.05$ and $\epsilon = 1$ minute.*

More simulation runs are needed for casualty patients, particularly at processes such as the PHARMACY, BLOOD TESTS, and X-RAYS, which are only visited by a small percentage of patients. The high number of simulations required to estimate waiting times in the DOCTORS queue indicates that there is a large amount of variance in the delays experienced by different patients in these queues.

6.1.2 Queue length

Plots of the queue length at each process in the 2015 and 2016 OPD set-ups are shown in Figures 6.1–6.6. The queue lengths have been smoothed over 5 minute intervals in order to make the graphs easier to read.

Stellenbosch University https://scholar.sun.ac.za

The queue lengths are presented in terms of the total number of patients in each queue, as well as the number of patients from each profile in the queue. In graphs depicting the overall queue length, the thin grey lines represent the queues in individual simulation runs and the thicker black line is the average of these results. In the second set of plots, the composition of each queue is illustrated in terms of the average number of patients from each profile.

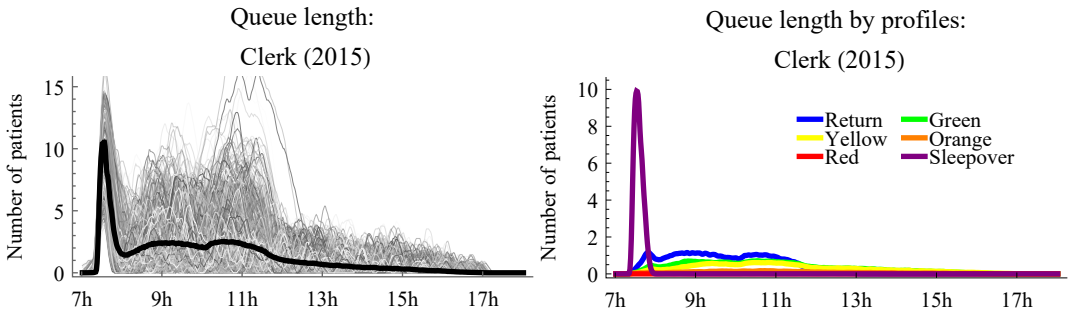


FIGURE 6.1: *Plots of the* CLERKS *queue length, based on the results of 2000 simulation runs.*

Figure 6.1 depicts the simulated queues for the CLERKS in the 2015 set-up. The length of this queue is closely linked to the arrival time distributions, since the CLERKS are the first process that patients go to when they arrive at the OPD. The 2016 results are not shown because there are no changes to the input data for the arrival distributions or the staff at this process.

In the simulation results the CLERKS queue tends to be longest at the beginning of the day due to sleepover and return patients who arrive before the staff begin working at 7h30. It decreases quickly between 7h30 and 8h00 and then stays in the range of about 1-5 patients for the rest of the morning. The afternoon queues are either empty or very short, as there are fewer new arrivals.

The graph on the right of Figure 6.1 shows the average number of patients from each profile in the CLERKS queue. The early morning queues are mostly sleepover patients, who are processed first and leave the queue quickly. The short queues during the rest of the morning are a mixture of return and casualty patients.

After the CLERKS, OPD patients proceed to the VITALS/TRIAGE queues. In the 2015 system there is a combined VITALS queue for all patient profiles, which is shown in the first column of graphs in Figure 6.2. The graphs in the second and third columns are the simulation results for the separate VITALS and TRIAGE queues in the 2016 system.

Like the CLERKS queues, the VITALS queues tend to have a peak at the beginning of the day when the sleepover patients are treated. This initial peak is similar in both the 2015 and 2016 VITALS queues, but their behaviour during the rest of the morning is different. The 2015 VITALS queues are more sensitive to the tea break between 10h00 and 11h00, which causes a peak in the average queue length at 11h00. The 2016 VITALS queues are not usually affected by the tea break because there are relatively few return patients arriving during this period.

The breakdown of the 2015 VITALS queue into different patient profiles is interesting, because the curves for the return, green, yellow and orange patients follow a similar pattern between 9h00 and 12h00. These similarities in shape are due to the fact that the VITALS queue is FCFS, so fluctuations in the queue length affect all of the different profiles.

In the breakdown of the 2016 VITALS queue the casualty patients are no longer present. This benefits the return patients, who experience fewer delays during the tea break, but does not change the way that the sleepover patients move through the VITALS queue.

Stellenbosch University https://scholar.sun.ac.za

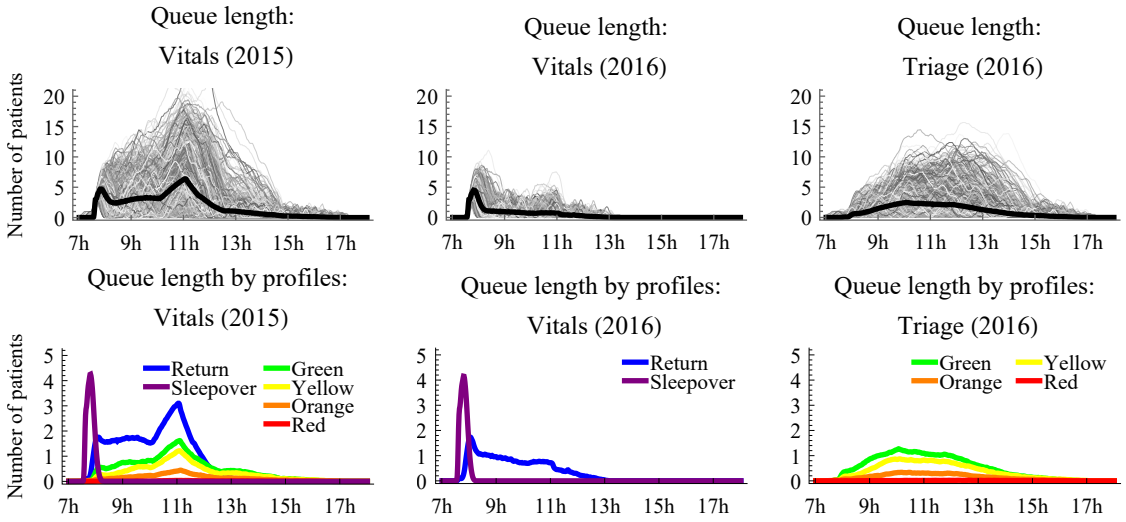


FIGURE 6.2: *Plots of the* VITALS/TRIAGE *queue length, based on the results of 2000 simulation runs.*

The 2016 TRIAGE queues on the right-hand side of Figure 6.2 follow a very different pattern to the VITALS queues. The TRIAGE queues do not have the same early peak, since there are fewer early morning casualty arrivals. Queues at this process only start to develop between 8h00 and 10h00, and generally stay below five patients. However, approximately 5% of the simulation runs resulted in queues of 10–15 patients. These unusually long queues occur because there is only one staff member allocated to the TRIAGE queue, which makes it difficult to process a large number of patients who arrive over a short space of time.

The separated VITALS/TRIAGE queues in the 2016 set-up are shorter and less prone to long backlogs than the combined 2015 VITALS queue. However, these differences are not a major improvement in terms of the overall efficiency of the system, since the 2015 queue was quite short to begin with. The main benefits of separating these queues in the 2016 system have more to do with the services provided at these processes, rather than the length of the queues. The new TRIAGE queue helps to organise casualty patients into categories and identify urgent patients who need faster treatment, which facilitates the priority discipline in the DOCTORS queues.

On average, the DOCTORS queues are the longest queues in both the 2015 and 2016 simulation results. The predicted queues for the 2015 set-up are shown in the first column of graphs in Figure 6.3 and the results for the separated casualty and return queues in the 2016 system are shown in the second and third columns.

The 2015 DOCTORS queue usually peaks during the lunch break, with an average of about 20 patients. The behaviour of this queue varies significantly on a daily basis, which is evident in the wide range of simulated queue lengths in Figure 6.3. In approximately 5% of these results the peak queue length is below 15 patients, while queues of more than 30 patients during the afternoon occurred in about 20% of the simulation runs. Regardless of the peak queue length, these queues grow steadily shorter after 14h00 because there are additional doctors available in the afternoon.

The graph of the different patient profiles in the 2015 DOCTORS queue indicates that the first patients to arrive at this queue in the morning are sleepover and return patients. The queue is dominated by return patients throughout the morning, which increases the delay for casualty patients. From about 13h00 this balance shifts, and the afternoon queues have a much higher

Stellenbosch University https://scholar.sun.ac.za

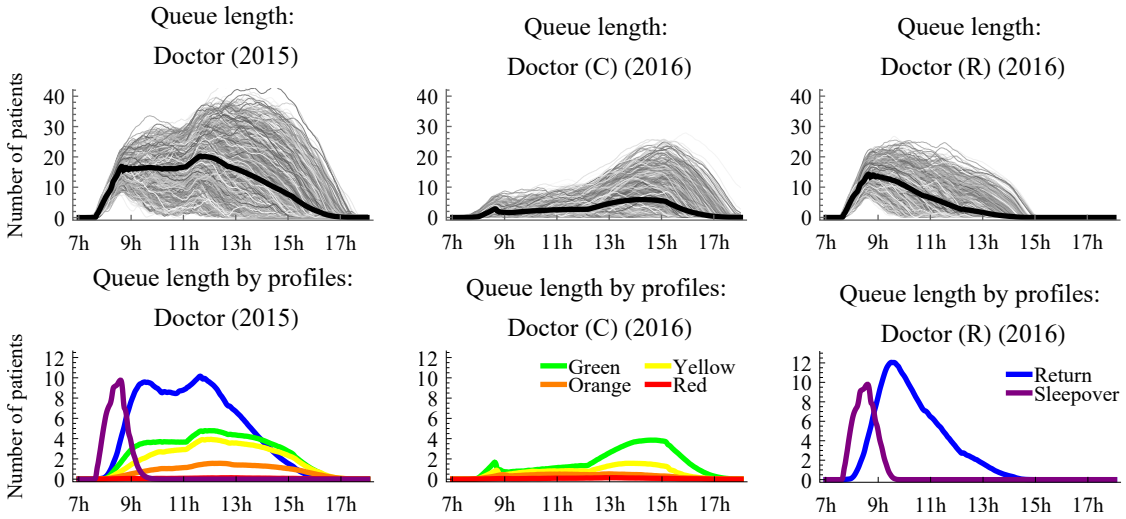


FIGURE 6.3: *Plots of the* DOCTORS *queue length, based on the results of 2000 simulation runs.*

proportion of casualty patients.

This pattern is problematic, since the long queue of return patients causes considerable delays for casualty patients who need urgent treatment. It also increases the number of sleepover patients, because casualty patients who only see a doctor in the afternoon are less likely to have the necessary BLOOD TESTS or X-RAYS completed on the same day.

The long DOCTORS queue also contributes to a general atmosphere of chaos and confusion in the OPD. There is no space in the OPD for a separate waiting room, so patients in this queue sit on benches along the walls of the main corridor through the OPD. This leads to a great deal of noise and congestion on busy days and makes it difficult for staff and patients to move freely around the OPD. The separated DOCTORS queues in the 2016 set-up are an improvement in this regard, as two shorter queues are much easier to manage than a single, long queue.

Figure 6.3 illustrates that the simulation results for the 2016 DOCTORS queue follow a very different pattern. The length of the return/sleepover queue peaks in the morning at around 9h00 and then decreases steadily over the course of the day. In about 3% of these results, the queue length remains as high as 15 patients until the early afternoon. The additional doctor available from 14h00 onwards is usually sufficient to deal with this backlog and all the patients in this queue are generally processed by about 15h00.

The simulation results for the CASUALTY DOCTORS queue tend to be spread more evenly across the day. On average, the queue length peaks at about 15h00 with 5–10 patients, although there are many instances where this peak did not occur at all. The peak length of this queue exceeded 10 patients in approximately 30% of the simulation runs, which indicates that very busy days can cause considerable delays.

The profile breakdown of the CASUALTY DOCTORS queue illustrates the effects of the priority queueing discipline. The afternoon queues consist mostly of green patients, even though these patients tend to arrive earlier in the day than other casualty patients. Yellow and orange patients benefit from higher priorities and are usually processed very efficiently in the morning and early afternoon.

The separation of the DOCTORS queue changes the general pattern of the BLOOD TESTS and X-RAYS queues. Figures 6.4 and 6.5 show that the simulation results for the 2016 parameters predict busier mornings at these processes than the corresponding results for the 2015 parameters. The

Stellenbosch University https://scholar.sun.ac.za

average morning queue length for BLOOD TESTS and X-RAYS increases slightly from 2015 to 2016, but this is not a serious problem because these processes have very short queues with an average length of fewer than 4 patients.

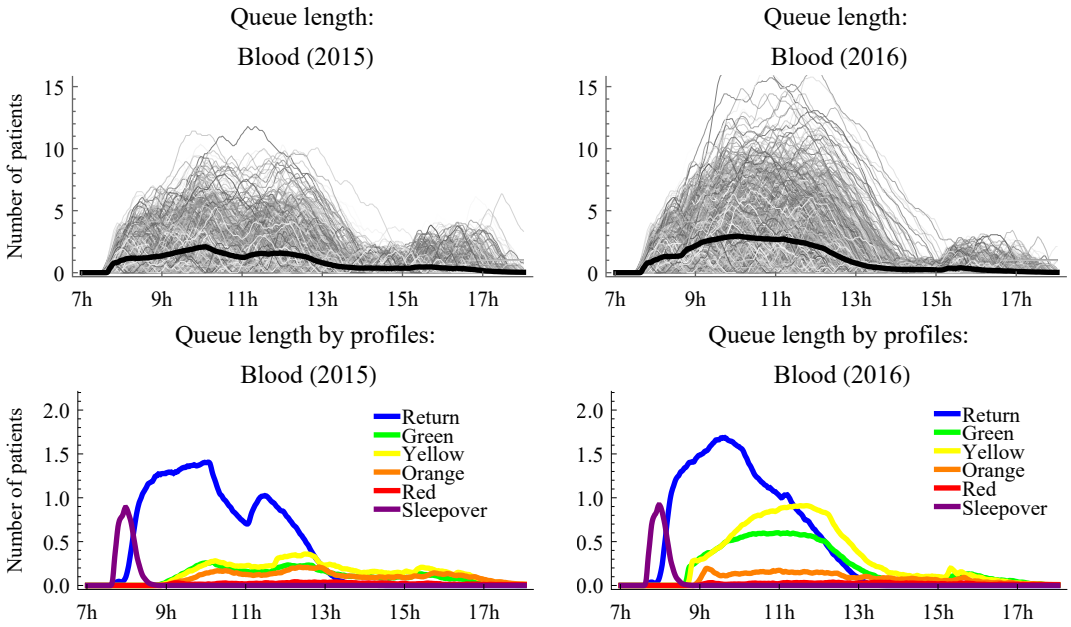


FIGURE 6.4: *Plots of the* BLOOD TESTS *queue length, based on the results of 2000 simulation runs.*

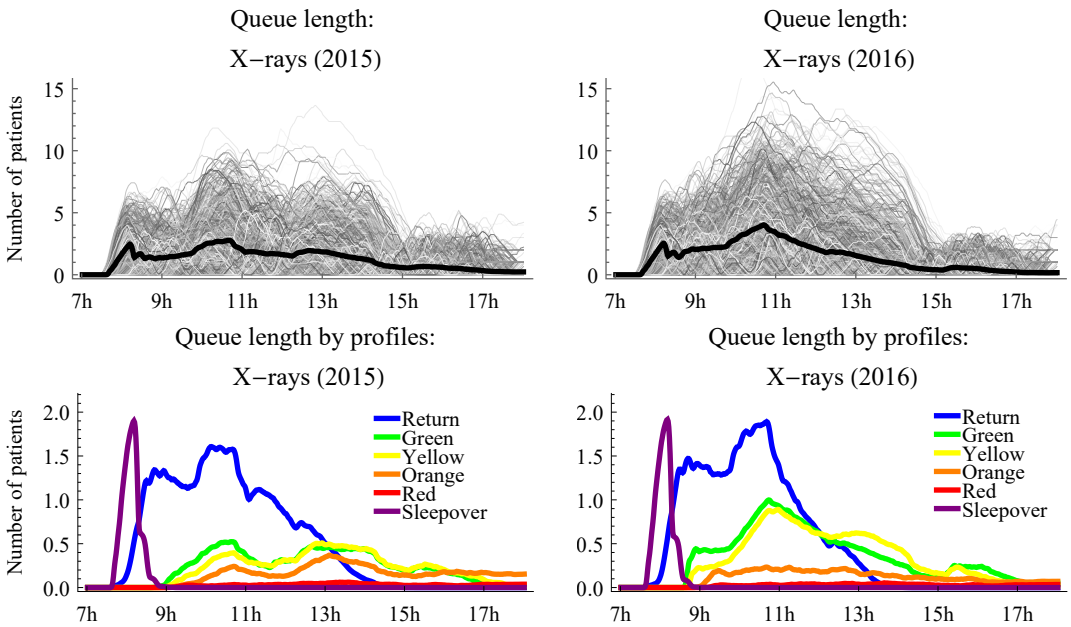


FIGURE 6.5: *Plots of the* X-RAYS *queue length, based on the results of 2000 simulation runs.*

A bigger concern is the increase in the frequency and duration of backlogs in the 2016 queues at these processes. The 2015 simulation results indicate that the BLOOD TESTS and X-RAYS queues only exceed 8 patients in 5% and 7% of the simulation runs, whereas about 17% of the 2016 simulation runs recorded queues of more than 8 patients at these processes. Queues of this length are likely to cause congestion and delays.

The profile breakdowns of the BLOOD TESTS and X-RAYS queues show that the longer queues in

Stellenbosch University https://scholar.sun.ac.za

the 2016 simulation runs are due to higher numbers of casualty patients at these processes during the morning and early afternoon. This pattern occurs because casualty patients are spending less time in the DOCTORS queues, which means that they get to the BLOOD TESTS/X-RAYS queues earlier.

Another factor that contributes to the number of casualty patients in these queues is the priority system in the 2016 set-up. At the BLOOD TESTS and X-RAYS queues, return patients have higher priority than yellow and green patients. This system helps to ensure that return patients have enough time to see the doctor and pick up their medication before the PHARMACY closes, but tends to cause delays for casualty patients in the mornings.

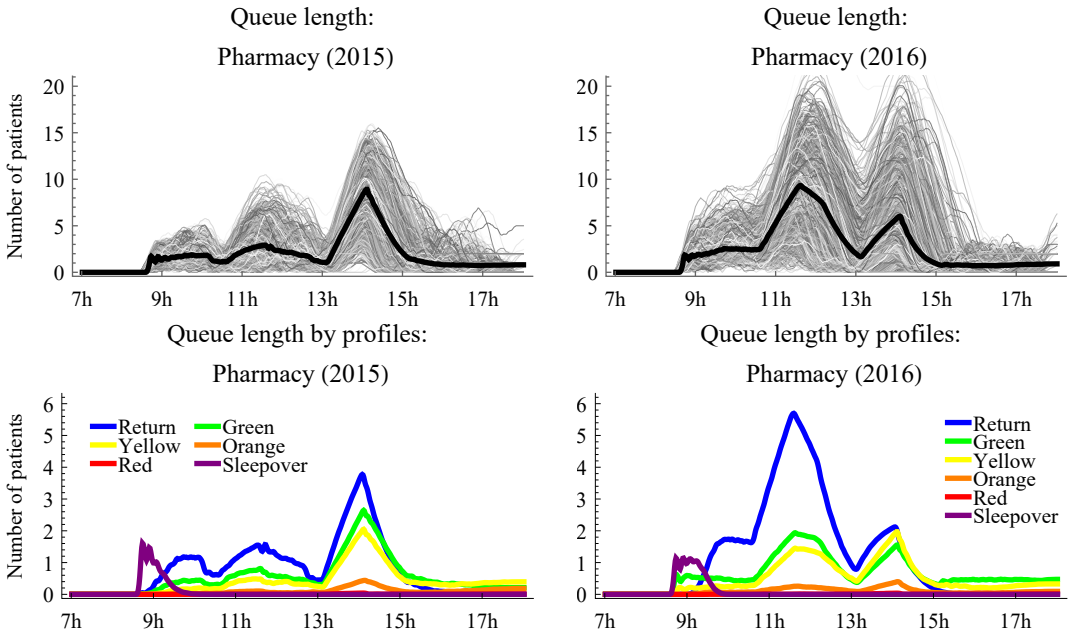


FIGURE 6.6: *Plots of the* PHARMACY *queue length, based on the results of 2000 simulation runs.*

The simulation results for the PHARMACY queues are shown in Figure 6.6. The overall queue length is sensitive to the number of staff on duty and peaks during the morning tea break and the lunch break. In the 2015 results, the morning peak is smaller due to the backlog in the DOCTORS queue. The 2016 simulation runs predict busier queues with higher peaks in the mornings, and slightly smaller queues during lunchtime.

There are slight differences in the composition of the 2015 and 2016 PHARMACY queues. In the 2016 results, green and yellow patients begin arriving at the PHARMACY earlier than return patients. From about 9h00 onwards, the number of return patients increases and causes longer queues during the morning tea breaks. Overall, the longer morning queues at the PHARMACY are a positive sign because they indicate that patients leave the OPD earlier in the 2016 simulations.

6.1.3 Waiting times

This section discusses the waiting times for different types of patients in the 2015 and 2016 simulation results. These waiting times are closely linked to the changes observed in the length and composition of each queue, particularly in queues where a priority discipline was implemented in the 2016 set-up.

The average waiting times for the different patient profiles are compared in Figure 6.7. In the

Stellenbosch University https://scholar.sun.ac.za

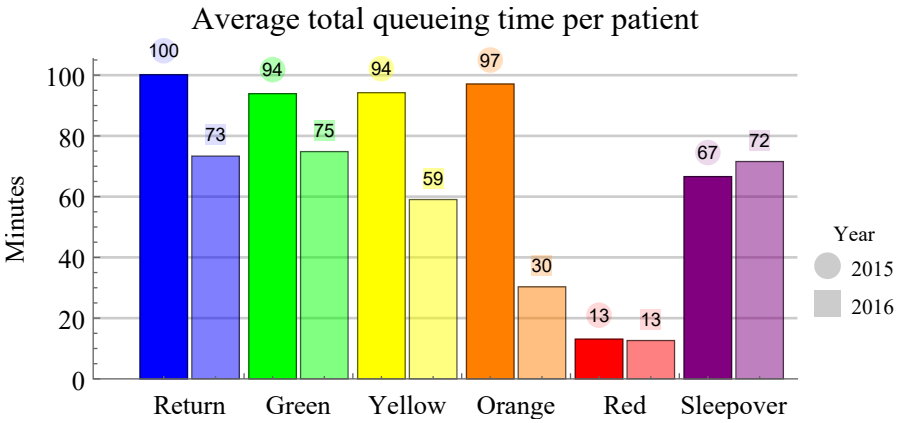


FIGURE 6.7: *A plot of the average total waiting time per patient in the 2015 and 2016 OPD set-ups, based on the results of 2000 simulation runs.*

2015 simulation results, the average waiting times for the first four profiles (return, green, yellow and orange) are all in the range of 94–100 minutes. Sleepover patients have a shorter average waiting time of about 67 minutes, and red patients have the shortest average waiting times (13 minutes).

The average waiting times in the 2016 results show significant improvements, especially for casualty patients. The average waiting times decrease by about 70% for orange patients, 37% for yellow patients and 20% for green patients. Waiting times for red patients do not decrease significantly, as these patients have the same priority in both set-ups. The 2016 results for the non-casualty patients are mixed: there is a 27% drop in the average waiting times for return patients, but the average waiting time for sleepover patients increases by 5 minutes.

The chart in Figure 6.8 illustrates some interesting changes to the range and spread of the waiting times observed in the 2015 and 2016 results. For return and orange patients, the maximum recorded waiting time decreases from 2015 to 2016 and the interquartile range is smaller. These results are positive, because they indicate a consistent improvement in waiting times and the fairness of the OPD queues.

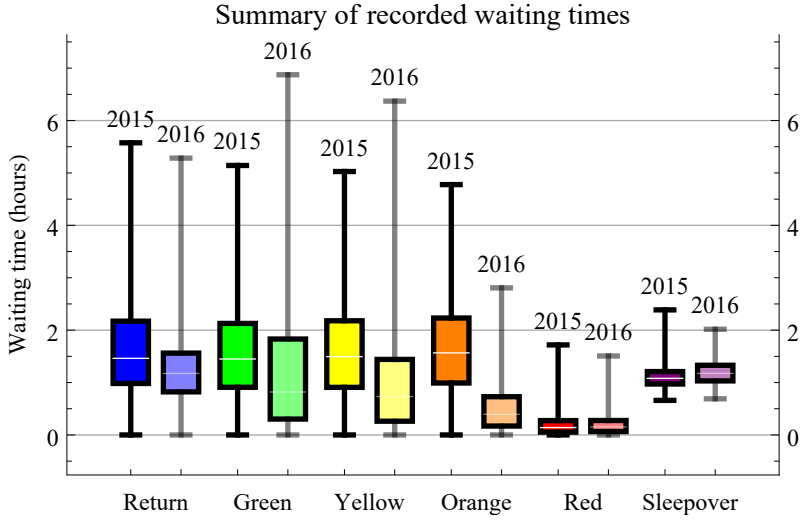


FIGURE 6.8: *Quartiles of the waiting times for different patient profiles, based on the results of 2000 simulation runs.*

Stellenbosch University https://scholar.sun.ac.za

The results for green and yellow patients are different: waiting times for both of these profiles have a higher maximum and a larger interquartile range in the 2016 results. Although there is an overall average improvement in the waiting times for these profiles, a wider range of waiting times means that the new system tends to be less fair to certain patients in these profiles.

To understand the source of the improved waiting times in the 2016 simulations, Figure 6.9 illustrates the contribution of each process to patient waiting times. As expected, these graphs show that the changes to the DOCTORS queue play the most significant role in lowering the waiting times for the majority of the OPD patients.

In the 2015 set-up, the DOCTORS queue is the biggest source of delays. Return, green, yellow and orange patients all spend an average of about 70 minutes in this queue, which accounts for 70–80% of their total average waiting time in the OPD.

The separation of the DOCTORS queue in the 2016 set-up results in a significant improvement for these profiles. On average, both green and return patients spend about 33 minutes less in the 2016 CASUALTY DOCTORS queue than in the combined queue in the 2015 set-up. The decrease in waiting times is sharper for orange and yellow patients — on average, orange patients spend only 8 minutes in the CASUALTY DOCTORS queue, and yellow patients experience an average delay of 17 minutes.

By contrast, red and sleepover patients do not benefit from the changes to the DOCTORS queue. The last two plots in Figure 6.9 show that these patients actually spend more time waiting to see a doctor in the 2016 results. This increase in the average waiting time is not an indication of the efficiency of the queues, since the 2016 CASUALTY DOCTORS/RETURN DOCTORS queues are generally shorter. Both of these profiles have certain advantages which lead to shorter waiting times in the 2015 DOCTORS queue, so it is understandable that they would gain less from the 2016 adjustments.

In the 2015 OPD set-up, red patients are already prioritised in the DOCTORS queue, so they are always seen by the first available doctor. Although this does not change in the 2016 system, the new CASUALTY DOCTORS queue has fewer staff than the combined queue in the 2015 set-up. The CASUALTY DOCTORS also tend to spend longer with each patient, so the amount of time that elapses between the arrival of a red patient and the availability of a doctor tends to be slightly longer.

A similar argument applies in the case of the sleepover patients. Even though they are not assigned a higher priority in the 2015 set-up, they are usually first in the queues because they are at the OPD before the other patients arrive. In the 2015 set-up they have access to a larger number of doctors early in the morning, so they are processed slightly faster.

In terms of the other OPD processes, the results in Figure 6.9 illustrate the effects of the changes in queue length that were discussed in the previous section. The longer BLOOD TESTS and X-RAYS queues result in higher average waiting times for yellow and green patients, who have the lowest priority at these processes. The high priority patients (orange and red) spend slightly less time in these queues, and return and sleepover patients do not experience any significant changes.

The differences in waiting times for the 2015 and 2016 set-ups were compared using paired t-tests and the Wilcoxon Signed Rank test. These tests are appropriate because the 2016 simulations were run using the same set of patients that were randomly generated in the 2015 simulations. The arrival times, treatment needs and treatment times for each patient were not changed, so any changes in waiting times are a direct result of the differences between the 2015 and 2016 set-ups.

Stellenbosch University https://scholar.sun.ac.za

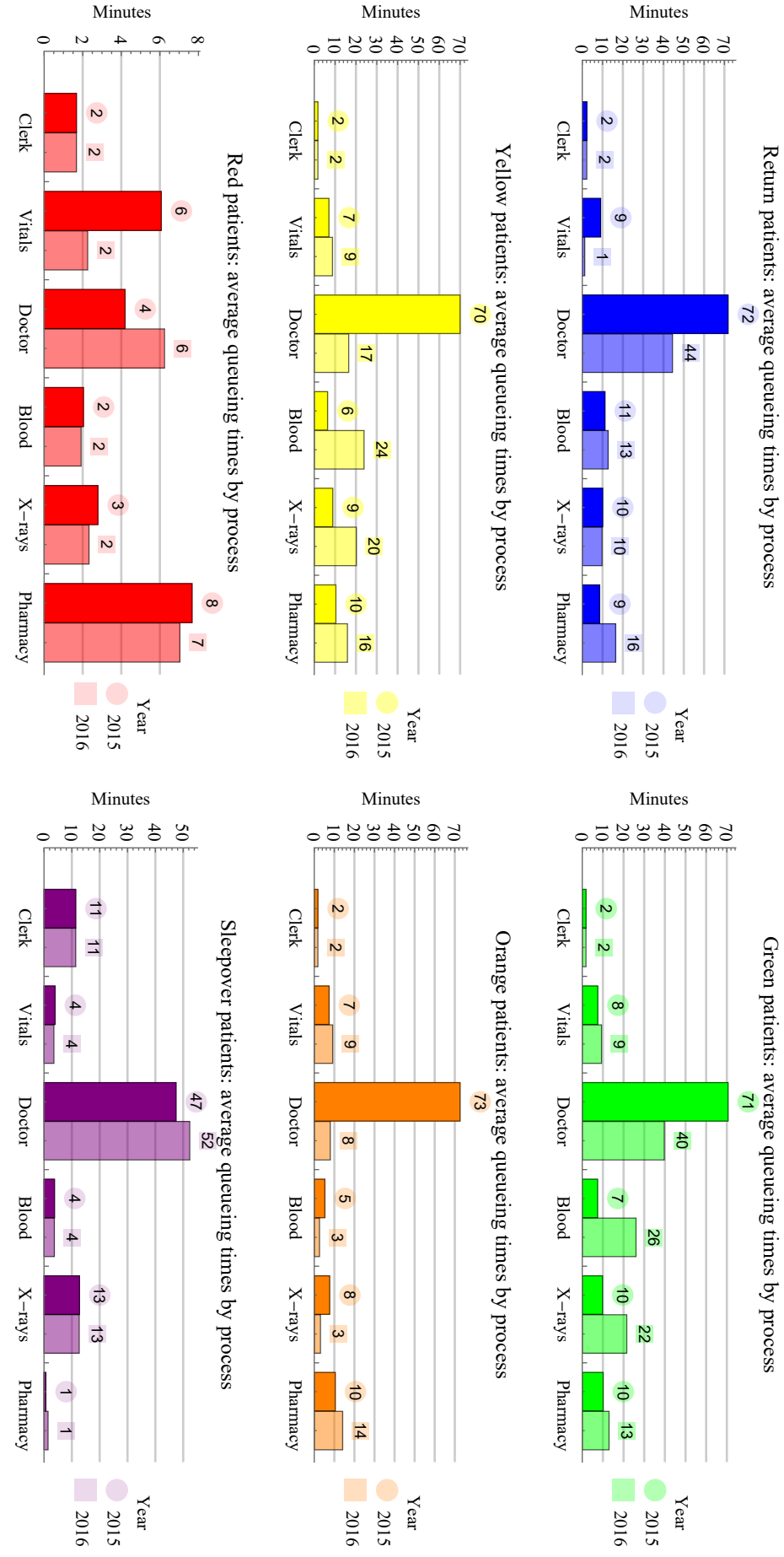


FIGURE 6.9: *The average waiting times at different processes in the 2015 and 2016 OPD set-ups, based on the results of 2000 simulation runs.*

Stellenbosch University https://scholar.sun.ac.za

The null hypothesis for each of these tests is

$$H_0 : \ \bar{w}_{i\ 2015}^{p} = \bar{w}_{i\ 2016}^{p} \ , \tag{6.3}$$

where $\bar{w}_i^p$ is the mean waiting time for patients from profile p at queue i. The waiting times at the CLERKS were excluded from this analysis, since there were no changes to this process and the results for the 2015 and 2016 set-ups are therefore identical.

	Return	Green	Yellow	Orange	Red	Sleepover
VITALS	$< 10^{-6}$ [1,2]	$< 10^{-6}$ [1,2]	$< 10^{-6}$ [1,2]	$< 10^{-6}$ [1,2]	$< 10^{-6}$ [1,2]	$< 10^{-6}$ [1,2]
DOCTORS	$< 10^{-6}$ [1,2]	$< 10^{-6}$ [1,2]	$< 10^{-6}$ [1,2]	$< 10^{-6}$ [1,2]	$< 10^{-6}$ [1,2]	$< 10^{-6}$ [1,2]
BLOOD TESTS	$< 10^{-6}$ [1,2]	$< 10^{-6}$ [1,2]	$< 10^{-6}$ [1,2]	$< 10^{-6}$ [1,2]	$0.1732^{[1]}$, $0.1314^{[2]}$	$0.944^{[1]}$, $0.0106^{[2]}$
X-RAYS	$< 10^{-6[1]}$, $0.001^{[2]}$	$< 10^{-6}$ [1,2]	$< 10^{-6}$ [1,2]	$< 10^{-6}$ [1,2]	$0.0003^{[1]}$, $0.0003^{[2]}$	$0.0018^{[1]}$, $0.0312^{[2]}$
PHARMACY	$< 10^{-6}$ [1,2]	$< 10^{-6}$ [1,2]	$< 10^{-6}$ [1,2]	$< 10^{-6}$ [1,2]	$0.2736^{[1]}$, $0.2444^{[2]}$	$< 10^{-6}$ [1,2]
Total	$< 10^{-6}$ [1,2]	$< 10^{-6}$ [1,2]	$< 10^{-6}$ [1,2]	$< 10^{-6}$ [1,2]	$0.6169^{[1]}$, $0.1252^{[2]}$	$< 10^{-6}$ [1,2]

[1]: Paired t-test, [2]: Wilcoxon Signed Rank test.

TABLE 6.2: *P-values for paired t-tests and the Wilcoxon Signed Rank test, comparing the waiting times for different patient profiles at each process in the 2015 and 2016 OPD set-ups. The null hypothesis that there is no difference in the mean waiting times is rejected for p-values smaller than 0.05.*

The p-values for these tests in Table 6.2 indicate that there was a significant change ($p < 0.05$) in average waiting times for return, green, yellow, orange and sleepover patients at all processes, as well as the total waiting times for these patients. There was also a significant change in the average waiting times for red patients in the VITALS, DOCTORS, and X-RAYS queues, but not in the BLOOD TESTS or PHARMACY queues.

The two tests reflect different outcomes for the average waiting times for sleepover patients in the BLOOD TESTS queue, although no underlying assumptions were violated in either test. Since the mean difference between these waiting times in the 2015 and 2016 simulations is less than 60 seconds, the statistical significance of this difference is not very important in terms of the overall efficiency of the two systems.

The final result discussed in this section is the number of waiting time targets that are missed in the 2015 and 2016 simulations. The waiting time targets are measured in terms of how long a patient has to wait between arriving at the OPD and seeing a doctor, not the overall waiting time for their visit. The graph in Figure 6.10 illustrates the percentage of patients in the simulation runs who were seen within the appropriate target time for their profile.

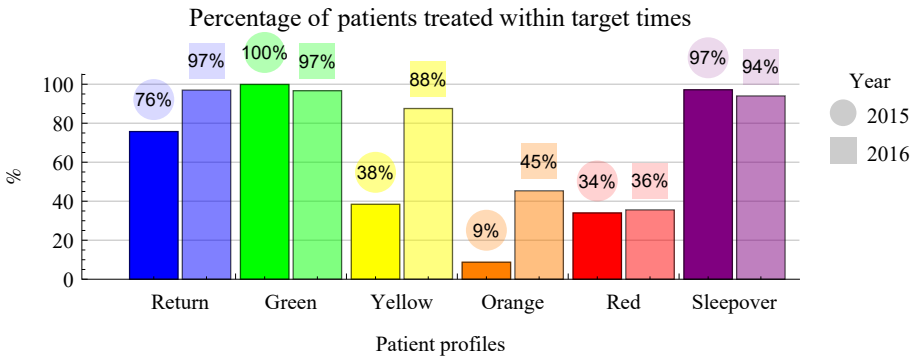


FIGURE 6.10: *A comparison of the number of patients treated within the target times in the 2015 and 2016 OPD set-ups, based on the results of 2000 simulation runs.*

Stellenbosch University https://scholar.sun.ac.za

The number of return patients who wait longer than two hours to see a doctor decreases from 24% to 3%, mostly due to the availability of additional doctors in the mornings. Although these patients are not as urgent as casualty patients, this improvement is still important. If these patients are processed in an efficient manner, they are more likely to continue following their treatments and return at the appropriate intervals.

In the case of the sleepover patients, overnight waiting time does not count towards their target waiting times, which are measured from 8h30 when the doctors begin working. Figure 6.10 indicates that nearly all of these patients see a doctor by 9h00 in both the 2015 simulations (97%) and the 2016 simulations (94 %).

The number of green patients not seen within the target time of 4 hours increases slightly to 3% in the 2016 results, which is due to the larger variance in waiting times for these patients. For the remaining casualty patients, the priority system in the 2016 set-up results in fewer missed targets. Yellow patients show the most significant improvement; only 12% of these patients wait longer than an hour in the 2016 set-up, compared to 62% in the 2015 set-up. The number of orange and red patients with missed targets drops from 91% to 55% and 66% to 64%.

The percentage of missed targets for the urgent casualty patients is relatively high when compared to other profiles, but is not a reflection of the overall efficiency of the OPD. The targets for these patients are very short — 10 minutes for orange patients and only 2 minutes for red patients — so they are frequently exceeded even when patients do not experience significant delays.

The overall decrease in patient waiting times and the number of missed targets indicates that most OPD patients are able to access treatments more quickly in the 2016 set-up. This is consistent with the changes observed in the length of the queues at each process, and also reflects how the priority queueing discipline leads to shorter delays for certain types of patients. Based on the simulation results, the 2016 OPD set-up is more efficient and better adapted to the needs of the different patient profiles.

6.1.4 Model verification and validation

The results of the OPD simulation model provide a great deal of information about the causes and effects of congestion in the queueing system. The purpose of the model is to provide insights that can be applied to the real-world problems, and it is therefore necessary to consider the accuracy and limitations of the simulation model and its results.

Verification

Verification tests are an important step in evaluating simulation models, although they are not directly concerned with the model's ability to produce realistic results. Rather, the purpose of verification tests is to ensure that **(a)** the high-level/conceptual model of the system has been implemented correctly in the simulation; and **(b)** the technical aspects of the simulation function properly (Banks *et al.*, 2004).

The OPD simulation model was verified through a close examination of the data generated in multiple simulation runs. This data was used to check various aspects of the model's implementation and logic, as outlined below.

1. The distributions of simulated patient counts, arrivals, and treatment needs/times were checked to ensure that they conformed to the input parameters.

Stellenbosch University https://scholar.sun.ac.za

2. Individual patient trajectories were examined to verify that **(a)** patients followed the correct path through the system; and **(b)** event times and waiting times were accurately recorded in the simulation results.

3. The simulated queues at each process were checked to confirm that **(a)** the appropriate queueing disciplines were followed; **(b)** queue lengths were accurately recorded in the simulation results; and **(c)** the composition of each queue matched the waiting times recorded for that queue.

4. The number of busy/available staff at each process was monitored to ensure that **(a)** the staff levels in the simulations matched the input data for staff schedules; **(b)** the staff break events were correctly implemented; and **(c)** the amount of time that staff were busy matched the treatment times for patients at that process.

Many of these checks were performed repeatedly during the process of implementing and debugging the simulation algorithm. They were also used to verify the final simulation model and the results presented in the previous sections. This analysis confirmed that the simulation model is an accurate implementation of the conceptual model outlined in Chapter 2.

Validation

The purpose of model validation is to evaluate both the accuracy and the usefulness of a model's results. Accuracy is generally determined by how well the model matches the real-world behaviour of the system, while usefulness is related to the model's ability to provide the information/insight required for the task at hand. Robinson (1999) suggests that model validity does not depend on any absolute standards, but rather whether the model is "sufficiently accurate" to fulfil its intended purpose.

The accuracy of the OPD simulation model could not be tested through a direct comparison with real-world data, since the OPD does not monitor or record the flow of patients through the different queues. However, several other validation methods were used, as explained below.

1. Face validation

   The simulation model was evaluated by the OPD clinical manager, Dr Ben Gaunt, as well as other members of the OPD staff. Based on their day-to-day experiences, staff were able to indicate whether the simulation results matched the observed trends in the OPD, such as **(a)** the normal/expected lengths of queues; **(b)** the peak queue lengths/busiest times at each process; **(c)** the flow of patients between different queues; and **(d)** the arrival/departure times of patients.

   Discrepancies between the simulation results and the observed trends were discussed with the OPD staff to identify elements of the model that could be improved. Feedback from these discussions resulted in numerous additions to the OPD model, such as time-dependent staff schedules, detailed treatment parameters, and priority queueing disciplines. The OPD model was evaluated, improved, and re-evaluated many times during the course of this research.

   Contact with the OPD staff also played an essential role in defining the purpose and scope of the model. These interactions helped to determine which aspects of the OPD system should be included in the model, as well as how to present the simulation results in a useful, informative manner.

Stellenbosch University https://scholar.sun.ac.za

2. Input-output validation

   An important aspect of the simulation model is its ability to illustrate the effect of changes to the OPD system. This was tested during discussions with hospital staff by varying the model's input parameters and confirming that the corresponding simulation results reflected the appropriate response. For example, increasing the number of staff at a process should result in shorter queues and waiting times at that process, while increasing the number of patients in the system should increase the general level of congestion.

3. Previous research

   Bertscher (2015) highlights several issues affecting patient flow in the OPD system, based on research conducted during December 2014. Although this study did not involve rigorous data collection or analysis, it does confirm some of the general trends in the simulation results for the 2015 OPD set-up. For example, key findings in Bertscher's report emphasise that

   (*a*) the DOCTORS queue is a bottleneck that limits the flow of patients through the OPD system;

   (*b*) the long delays that patients experience in this queue are the main source of inefficiency in the OPD;

   (*c*) fluctuations in patient arrivals have a significant effect on congestion — the OPD functions well when there are relatively few patients, but the system is not equipped to handle large numbers of patients efficiently.

   Bertscher (2015) also mentions other sources of inefficiency in the OPD which are not included in the simulation model. Examples of these problems include

   (*d*) routing confusion — some patients may not follow the correct route through the OPD because they do not understand the system or receive incorrect instructions from staff;

   (*e*) missing equipment/supplies — staff (particularly doctors) spend too much time searching for medical equipment and supplies that have been removed from their consultation rooms.

   These observations illustrate some of the limitations of the simulation model, which does not reflect the time that staff spend on unforeseen tasks. There are several other sources of inefficiency that the model does not account for, such as equipment failures, missing patient records, and staff members arriving late for their shift or missing work.

   These limitations should be taken into account when analysing the simulation model's results, which reflect the flow of patients through the OPD in the absence of these types of delays and interruptions. In reality, patient's waiting times may be longer and more varied than the waiting times in the simulation results, and there may also be greater variance in the length of the OPD queues. The simulation model's results should therefore not be interpreted as a direct measure of the actual patient waiting times in the OPD.

Based on the information gathered from these sources, the simulation model is sufficiently accurate to fulfil its intended purpose. It can be used to develop a better understanding of the OPD system, including the behaviour of the OPD queues and the experiences of different patient profiles. The model can also provide insight into the causes and effects of congestion and the impact of changes to the system.

Stellenbosch University https://scholar.sun.ac.za

6.1.5 Sensitivity analysis

The sensitivity analysis in this section focusses on the treatment time parameters in the OPD simulation model. Many of these parameters are based on discussions with the OPD staff and data recorded from short observations of the OPD queues, so it is quite likely that these parameters are over- or under-estimated. The purpose of this section is to illustrate how these inaccuracies might affect the simulation results.

The sensitivity analysis for these parameters was performed by increasing and decreasing the treatment times at a specific process in the simulation model by 10%, 20% and 50%. Changes of 10–20% are within the range of treatments times that might occur at these processes. Changes of 50% are not realistic, but have been included in the sensitivity analysis because they demonstrate the impact of the treatment time parameters more effectively than the smaller changes.

The treatment parameter values were tested in the 2015 set-up using the same simulation data generated for the results in § 6.1.2. No changes were made to the arrival times of these patients, and their original treatment times were also kept fixed at all processes other than the process where higher or lower treatment times were considered. Changes to the treatment times at this process were implemented by increasing or decreasing the treatment times generated in the original simulations by the appropriate percentage.

Waiting times

Table 6.4 gives an overview of the change in the average waiting times for each patient profile when the treatment parameters for a single process are increased or decreased. These results indicate that the treatment times for the DOCTORS queue have a much greater impact on patients' total waiting times than any of the other processes. The smallest adjustment that was tested, a 10% change in these parameters, increased the average waiting time for all patients by 19% (18 minutes) and decreased waiting times by 18% (16 minutes) when lower treatment times were used. A 20% change in these parameters increased and decreased average waiting times by 38% and 33%, while a 50% change in these treatment times resulted in average waiting times 50% lower and 96% higher than the original simulations.

The changes in waiting times are similar for return, green, yellow and orange patients. However, red patients and sleepover patients are affected less than the other profiles, since these patients do not have to wait in the DOCTORS queue during the busy periods.

Changes to treatment times at other processes did not generally cause a change of more than 5% in the average total waiting times. Two exceptions to this rule are the BLOOD TESTS and X-RAYS queues, where a 50% increase in treatment times results in an increase of 7% in the average total waiting times for all patients. Red, orange and yellow patients are most affected by these increases, since a higher proportion of these patients are required to stand in these queues. The average amount of time that patients spend in the OPD increases by approximately 10 minutes for yellow patients, 23 minutes for orange patients and 4 minutes for red patients.

Total waiting times also increase significantly when treatment times at the PHARMACY are increased by 50%. In this case, return, green and yellow patients are most affected, experiencing increases of 24%, 30% and 29% in their average total waiting times. These profiles make up the majority of the OPD patients, so the corresponding increase in the average waiting times is 28 minutes.

The results in Table 6.4 provide some insight into why the PHARMACY parameters have a bigger effect on waiting times than the parameters at other processes. As expected, the table shows that

Stellenbosch University https://scholar.sun.ac.za

	Percentage change in average total waiting times.													
Parameters changed	Return		Green		Yellow		Orange		Red		Sleepover		All patients	
	−	+	−	+	−	+	−	+	−	+	−	+	−	+
CLERKS														
10%	1%	-1%	1%	-1%	1%	-1%	1%	-1%	3%	2%	1%	-1%	1%	-1%
20%	1%	-1%	1%	-1%	1%	-1%	2%	-1%	3%	-1%	2%	-2%	1%	-1%
50%	3%	-3%	3%	-3%	4%	-3%	4%	-5%	2%	0%	4%	-5%	3%	-3%
VITALS														
10%	1%	-1%	0%	0%	0%	0%	0%	0%	-7%	19%	1%	-1%	1%	-1%
20%	2%	-3%	1%	-1%	1%	-1%	1%	0%	-14%	53%	3%	-3%	1%	-2%
50%	4%	-5%	2%	2%	2%	2%	3%	3%	-24%	237%	6%	-6%	3%	-1%
DOCTORS														
10%	-18%	18%	-20%	22%	-20%	23%	-20%	24%	2%	5%	-2%	2%	-18%	19%
20%	-34%	36%	-38%	44%	-37%	47%	-35%	50%	1%	6%	-4%	5%	-33%	38%
50%	-51%	90%	-58%	111%	-55%	119%	-51%	120%	-5%	13%	-9%	12%	-50%	96%
BLOOD TESTS														
10%	-1%	1%	0%	0%	0%	1%	-1%	2%	0%	5%	0%	0%	0%	1%
20%	-1%	2%	0%	0%	-1%	2%	-2%	5%	-5%	12%	0%	0%	-1%	2%
50%	-2%	7%	1%	2%	0%	10%	-2%	24%	-9%	29%	1%	0%	-1%	7%
X-RAYS														
10%	0%	1%	0%	0%	0%	1%	-1%	2%	-4%	9%	0%	0%	0%	1%
20%	-1%	2%	0%	1%	-1%	2%	-2%	5%	-5%	14%	0%	0%	-1%	2%
50%	-1%	5%	-1%	6%	-1%	11%	-3%	24%	-11%	33%	1%	0%	-1%	7%
PHARMACY														
10%	-1%	2%	-2%	3%	-1%	2%	-1%	1%	0%	1%	0%	0%	-1%	2%
20%	-2%	6%	-3%	7%	-2%	6%	-1%	2%	-1%	1%	0%	1%	-2%	6%
50%	-4%	30%	-5%	38%	-4%	37%	-2%	18%	-1%	4%	-1%	2%	-4%	30%

	Change in average total waiting times (minutes).													
Parameters changed	Return		Green		Yellow		Orange		Red		Sleepover		All patients	
	−	+	−	+	−	+	−	+	−	+	−	+	−	+
CLERKS														
10%	0.7	-0.5	0.6	-0.5	0.6	-0.6	0.9	-0.6	0.4	0.3	0.6	-0.5	0.7	-0.5
20%	1.3	-1.1	1.3	-1.2	1.3	-1.2	1.7	-1.4	0.3	-0.1	1.1	-1.1	1.3	-1.1
50%	3.1	-3.2	3.2	-2.4	3.5	-2.5	3.7	-4.6	0.2	0	2.8	-3.1	3.1	-2.9
VITALS														
10%	1.2	-1.3	0.2	-0.2	0.3	-0.3	0.5	-0.1	-0.9	2.4	0.9	-0.9	0.7	-0.7
20%	2	-2.7	0.6	-0.6	0.7	-0.5	0.9	-0.5	-1.7	6.7	1.8	-1.9	1.3	-1.5
50%	4.2	-4.8	2.1	1.8	2.2	2.2	2.4	3	-3	30	3.7	-4.2	3.1	-1.1
DOCTORS														
10%	-18	18	-19	20	-19	22	-20	23	0.2	0.6	-1.4	1.5	-16	18
20%	-34	36	-35	41	-35	44	-34	48	0.1	0.8	-2.8	3	-30	35
50%	-51	90	-54	105	-52	112	-49	116	-0.7	1.6	-6	7.8	-46	88
BLOOD TESTS														
10%	-0.7	1.1	0.2	0	-0.2	0.8	-0.9	1.8	0	0.6	0.1	-0.1	-0.3	0.7
20%	-1.3	2.4	0.2	0.1	-0.5	2	-1.6	4.4	-0.6	1.5	0.2	-0.1	-0.6	1.6
50%	-2.1	6.8	0.7	1.8	-0.5	9.8	-2.3	23	-1.2	3.7	0.7	-0.1	-0.9	6.3
X-RAYS														
10%	-0.4	0.7	-0.1	0.3	-0.4	0.8	-1.2	1.8	-0.5	1.1	0.1	0	-0.3	0.6
20%	-0.8	1.5	-0.2	0.8	-0.7	1.8	-1.8	4.6	-0.7	1.7	0.1	-0.1	-0.6	1.4
50%	-1.1	5.4	-0.6	5.5	-1.3	10	-3.4	23	-1.4	4.2	0.3	-0.1	-1	6.6
PHARMACY														
10%	-1.4	2.2	-1.5	2.4	-1.3	2.1	-0.6	0.8	0	0.1	-0.1	0.2	-1.2	1.9
20%	-2.5	5.8	-2.6	6.5	-2.3	5.8	-1	2.3	-0.1	0.1	-0.2	0.4	-2.1	5.1
50%	-4.2	30	-4.6	36	-4.1	35	-1.9	17	-0.1	0.6	-0.4	1.5	-3.7	28

TABLE 6.3: *The change in average total waiting times for each profile due to increases and decreases in the treatment time parameters.*

Stellenbosch University https://scholar.sun.ac.za

Percentage change in average waiting times at each process.														
Parameters changed	CLERKS		VITALS		DOCTORS		BLOOD TESTS		X-RAYS		PHARMACY		Total	
	−	+	−	+	−	+	−	+	−	+	−	+	−	+
CLERKS														
10%	-22%	31%	8%	-10%	1%	-1%	0%	0%	0%	1%	0%	0%	1%	-1%
20%	-38%	75%	13%	-24%	2%	-3%	0%	0%	0%	1%	0%	0%	1%	-1%
50%	-63%	339%	23%	-59%	5%	-13%	0%	-4%	0%	2%	0%	-1%	3%	-3%
VITALS														
10%	-	-	-41%	65%	5%	-8%	5%	-6%	1%	1%	1%	-1%	1%	-1%
20%	-	-	-67%	159%	9%	-20%	8%	-12%	1%	3%	0%	-2%	1%	-2%
50%	-	-	-96%	602%	15%	-67%	10%	-28%	5%	-4%	1%	-11%	3%	-1%
DOCTORS														
10%	-	-	-	-	-28%	28%	16%	-10%	21%	-11%	17%	-14%	-18%	19%
20%	-	-	-	-	-53%	56%	42%	-17%	48%	-13%	33%	-23%	-33%	38%
50%	-	-	-	-	-87%	136%	103%	-34%	78%	-11%	90%	-31%	-50%	96%
BLOOD TESTS														
10%	-	-	-	-	1%	-1%	-32%	46%	1%	-1%	-1%	2%	0%	1%
20%	-	-	-	-	2%	-2%	-55%	107%	1%	-4%	-2%	2%	-1%	2%
50%	-	-	-	-	3%	-7%	-89%	403%	2%	-13%	-4%	-3%	-1%	7%
X-RAYS														
10%	-	-	-	-	1%	-1%	0%	0%	-28%	39%	0%	-1%	0%	1%
20%	-	-	-	-	2%	-2%	-1%	1%	-47%	92%	0%	-4%	-1%	2%
50%	-	-	-	-	3%	-7%	-1%	3%	-81%	353%	-4%	-13%	-1%	7%
PHARMACY														
10%	-	-	-	-	-	-	-	-	-	-	-16%	25%	-1%	2%
20%	-	-	-	-	-	-	-	-	-	-	-28%	67%	-2%	6%
50%	-	-	-	-	-	-	-	-	-	-	-48%	366%	-4%	30%

Change in average waiting times at each process (minutes).														
Parameters changed	CLERKS		VITALS		DOCTORS		BLOOD TESTS		X-RAYS		PHARMACY		Total	
	−	+	−	+	−	+	−	+	−	+	−	+	−	+
CLERKS														
10%	-0.7	1	0.7	-0.9	0.7	-0.8	0	0	0	0.1	0	0	0.7	-0.5
20%	-1.2	2.4	1.1	-2	1.4	-1.7	0	0	0	0.1	0	0	1.3	-1.1
50%	-2	11	2	-5.1	3.4	-8.9	0	-0.4	0	0.2	0	-0.1	3.1	-2.9
VITALS														
10%	-	-	-3.6	5.6	3.6	-5.5	0.4	-0.5	0.1	0.1	0.1	-0.1	0.7	-0.7
20%	-	-	-5.8	14	6.1	-13	0.7	-1	0.1	0.3	0	-0.1	1.3	-1.5
50%	-	-	-8.3	52	9.9	-45	0.8	-2.3	0.4	-0.4	0.1	-0.8	3.1	-1.1
DOCTORS														
10%	-	-	-	-	-19	19	1.3	-0.8	2	-1	1.3	-1	-16	18
20%	-	-	-	-	-36	38	3.4	-1.4	4.5	-1.2	2.5	-1.7	-30	35
50%	-	-	-	-	-59	92	8.5	-2.8	7.3	-1	6.9	-2.3	-46	88
BLOOD TESTS														
10%	-	-	-	-	0.6	-0.7	-2.6	3.8	0.1	-0.1	-0.1	0.1	-0.3	0.7
20%	-	-	-	-	1.1	-1.6	-4.5	8.8	0.1	-0.3	-0.2	0.2	-0.6	1.6
50%	-	-	-	-	2	-4.9	-7.3	33	0.2	-1.2	-0.3	-0.2	-0.9	6.3
X-RAYS														
10%	-	-	-	-	0.6	-0.7	0	0	-2.6	3.7	0	-0.1	-0.3	0.6
20%	-	-	-	-	1.1	-1.6	-0.1	0.1	-4.4	8.6	0	-0.3	-0.6	1.4
50%	-	-	-	-	2.3	-5	-0.1	0.2	-7.6	33	-0.3	-1	-1	6.6
PHARMACY														
10%	-	-	-	-	-	-	-	-	-	-	-1.2	1.9	-1.2	1.9
20%	-	-	-	-	-	-	-	-	-	-	-2.1	5.1	-2.1	5.1
50%	-	-	-	-	-	-	-	-	-	-	-3.7	28	-3.7	28

TABLE 6.4: *The change in average waiting times (minutes) at each process due to increases and decreases in the treatment time parameters.*

Stellenbosch University https://scholar.sun.ac.za

decreasing/increasing treatment times at a particular process decreases/increases the amount of time that patients spend in the queue for that process. However, these changes tend to be balanced out at the next process, and so they do not have a very large effect on the total waiting times. Since the PHARMACY is always the last queue that a patient will stand in before leaving the OPD, the additional delays that are experienced in this queue cannot be balanced by shorter waiting times at other processes in the system.

This relationship between the waiting times at each successive process in the network also has implications for the CLERKS and VITALS processes, which are the first two queues in the network. Changes to the treatment time parameters in these queues lead to unexpected results — shorter treatment times at these processes increase the average total waiting time, while longer treatment times decrease the average total waiting time.

These results appear to be contradictory, since they suggest that additional delays at these processes actually increase the overall efficiency of the network. To understand why this happens, it is necessary to consider the routing of different patient profiles through the OPD. In the 2015 set-up, casualty patients go directly to the DOCTORS queue after VITALS, while return and sleepover patients may have BLOOD TESTS or X-RAYS before they see the doctor.

When the CLERKS or VITALS queues are treated more efficiently, casualty patients have a slight advantage in the DOCTORS queue, because they can join the queue before many of the return patients have completed their BLOOD TESTS and X-RAYS. This means that there will be a larger number of casualty patients ahead of return patients in the DOCTORS queue. Since casualty patients have longer treatment times, a small number of casualty patients near the front of this queue can cause additional delays for a much larger group of return patients.

The same explanation can be applied to increases in the treatment times at the CLERKS or VITALS queues. Longer delays at these processes mean that there are fewer casualty patients joining the DOCTORS queue while return patients are waiting for tests, so return patients are less likely to be delayed by time-consuming casualty cases. The return patients are the largest profile in the OPD model, so the average total waiting times are strongly influenced by changes in waiting times for this profile.

Queue length

Figures 6.11–6.16 illustrate how the length of the different OPD queues is influenced by changes to the treatment parameters. The left-hand graphs show the average length of the queues at different times in the day, while the right-hand graphs indicate the cumulative number of patients treated at each process over the course of the day.

These graphs reflect a similar pattern to the changes observed in the average waiting times at each process (Table 6.4). The queue length graphs show that decreases/increases in the treatment time parameters at a particular process cause a decrease/increase in the average queue length at that process, and a corresponding increase/decrease in queue length at the next process in the network.

The cumulative treatment graphs show significantly less variability for different treatment time parameters, which indicates that the changes in the average queue lengths do not necessarily influence how quickly patients progress through the OPD. In fact, the treatment times at the doctors are the only parameters which do have a noticeable effect on the overall flow of patients through the system.

When the treatment times at the DOCTORS queue are decreased, Figure 6.13 shows that a

Stellenbosch University https://scholar.sun.ac.za

greater number of patients are treated during the morning and early afternoon. Figures 6.14–6.15 indicate that the improved efficiency of the DOCTORS queue allows casualty patients to join the BLOOD TESTS and X-RAYS queues earlier in the day, so the cumulative number of treatments at these processes increases. Most importantly, the number of patients leaving the PHARMACY during the morning and early afternoon is also higher (Figure 6.16), which shows that the total amount of time that patients are spending in the OPD has decreased.

When the treatment time parameters for the DOCTORS queue are increased, the reverse is true. The cumulative number of treatments at the DOCTORS queue is lower (Figure 6.13), and patients receive BLOOD TESTS and X-RAYS later in the day (Figures 6.14–6.15). These delays extend to the PHARMACY queue, where Figure 6.16 indicates that longer treatment times at the DOCTORS queue result in patients leaving the OPD later in the day. When the treatment parameters for the DOCTORS queue are increased by 50%, there are an average of 9 patients still queueing for treatments when the OPD closes.

Based on these results, the treatment times for the DOCTORS queue are the most influential parameters in the simulation model, and it is therefore important to have accurate estimates for these parameters. Since detailed data regarding the treatment times for different types of patients at this process is not available, this information should be prioritised during future data collection projects.

The treatment parameters at processes other than the DOCTORS queue are less influential. Changes to these parameters can result in longer or shorter queues at particular processes, but the effects of these changes are usually absorbed by correspondingly shorter or longer queues at the next process in the network. Over- or under-estimated treatment times at these processes are unlikely to distort the overall behaviour of the system, since the flow of patients through the OPD tends to be restricted by the long delays in the DOCTORS queue.

Stellenbosch University https://scholar.sun.ac.za

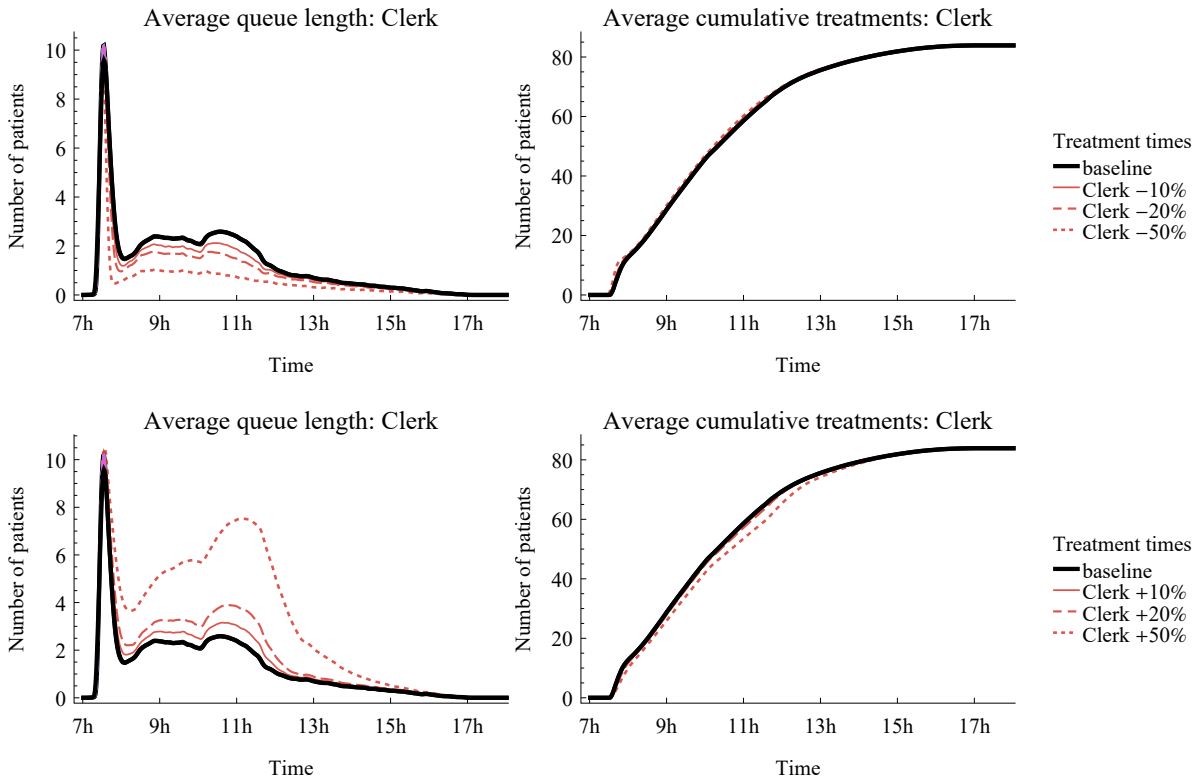


FIGURE 6.11: *The effect of changes to treatment time parameters on the* CLERKS *queue.*

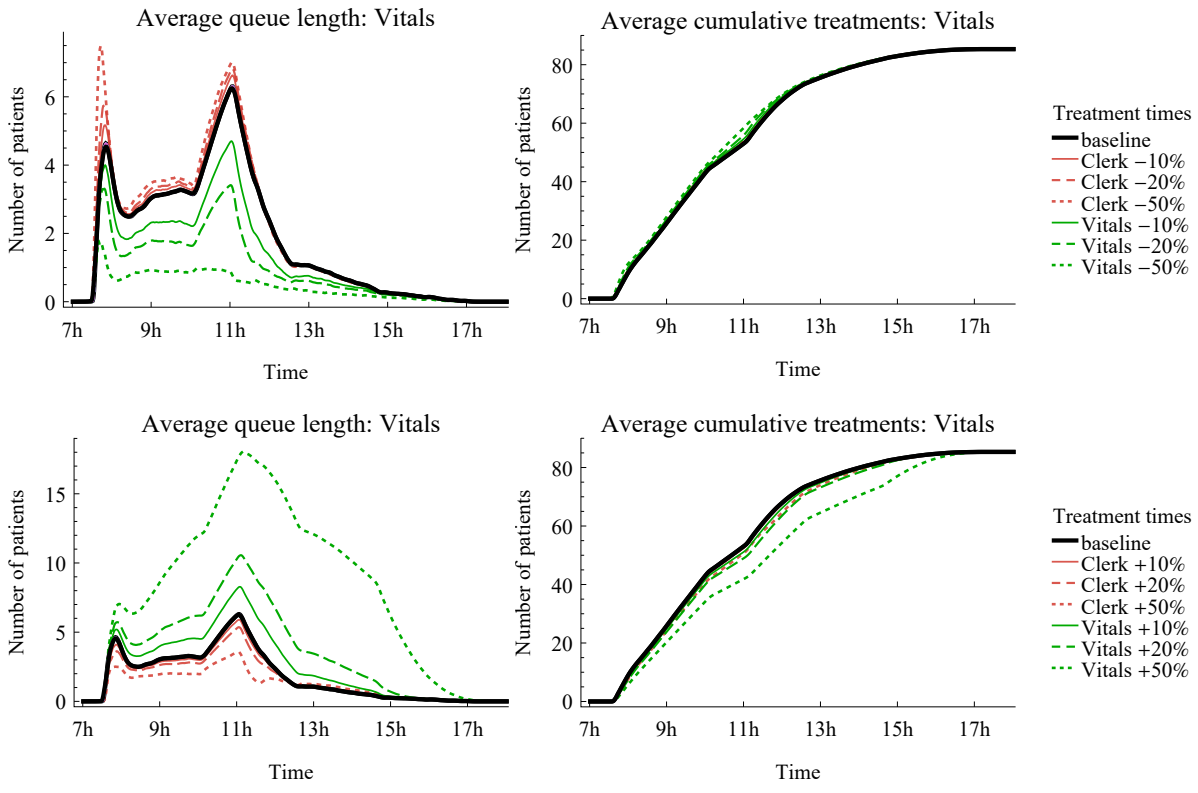


FIGURE 6.12: *The effect of changes to treatment time parameters on the* VITALS *queue.*

Stellenbosch University https://scholar.sun.ac.za

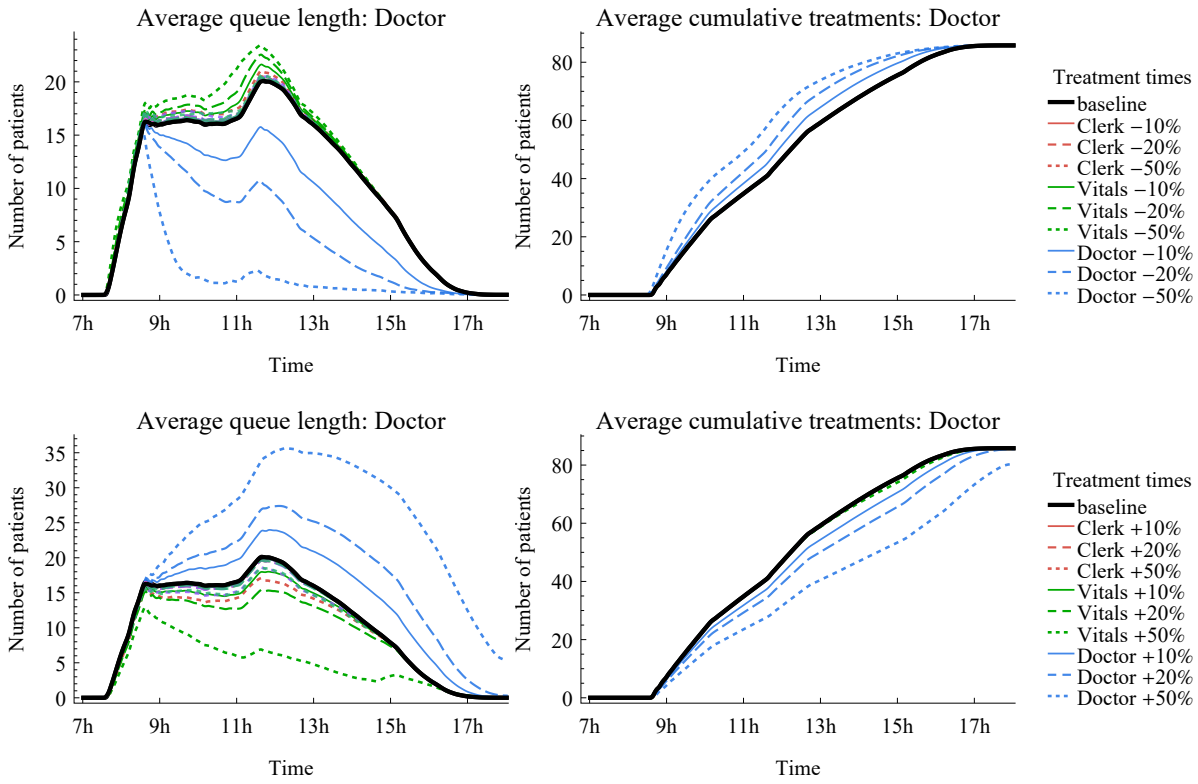


FIGURE 6.13: *The effect of changes to treatment time parameters on the* DOCTORS *queue.*

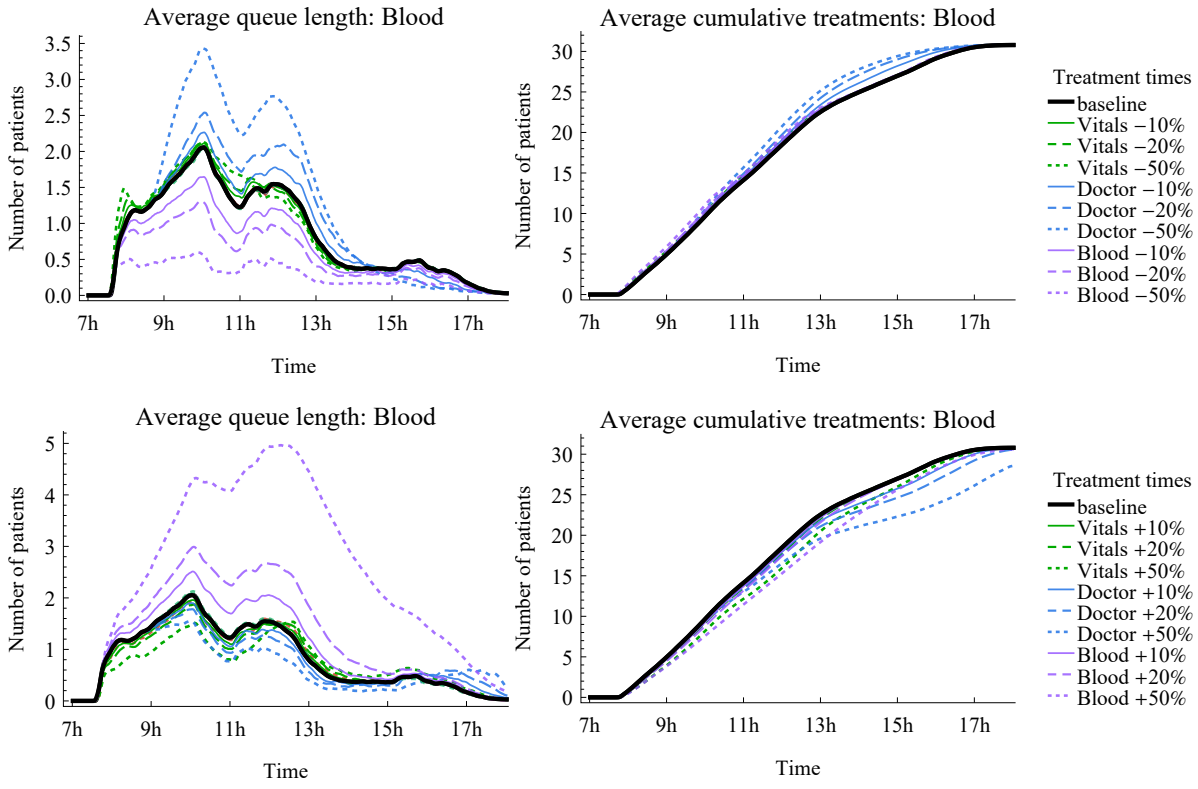


FIGURE 6.14: *The effect of changes to treatment time parameters on the* BLOOD TESTS *queue.*

Stellenbosch University https://scholar.sun.ac.za

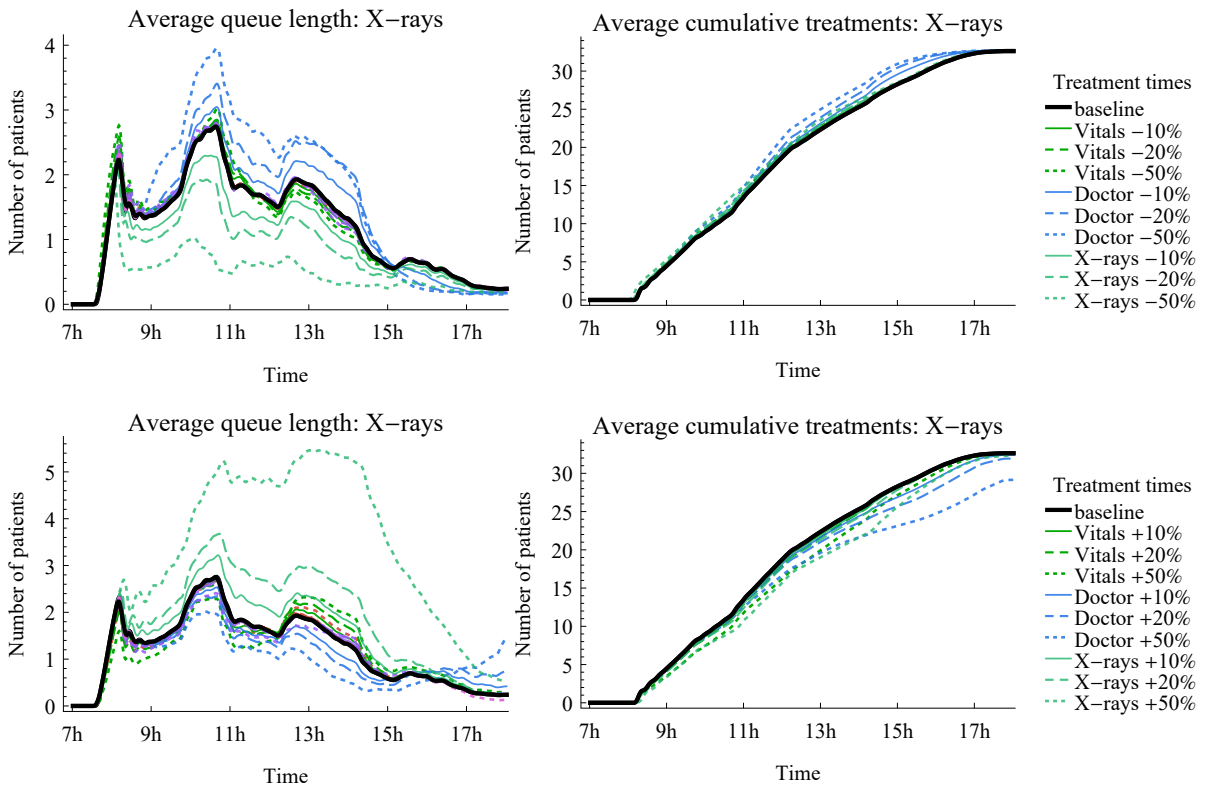


FIGURE 6.15: *The effect of changes to treatment time parameters on the* X-RAYS *queue.*

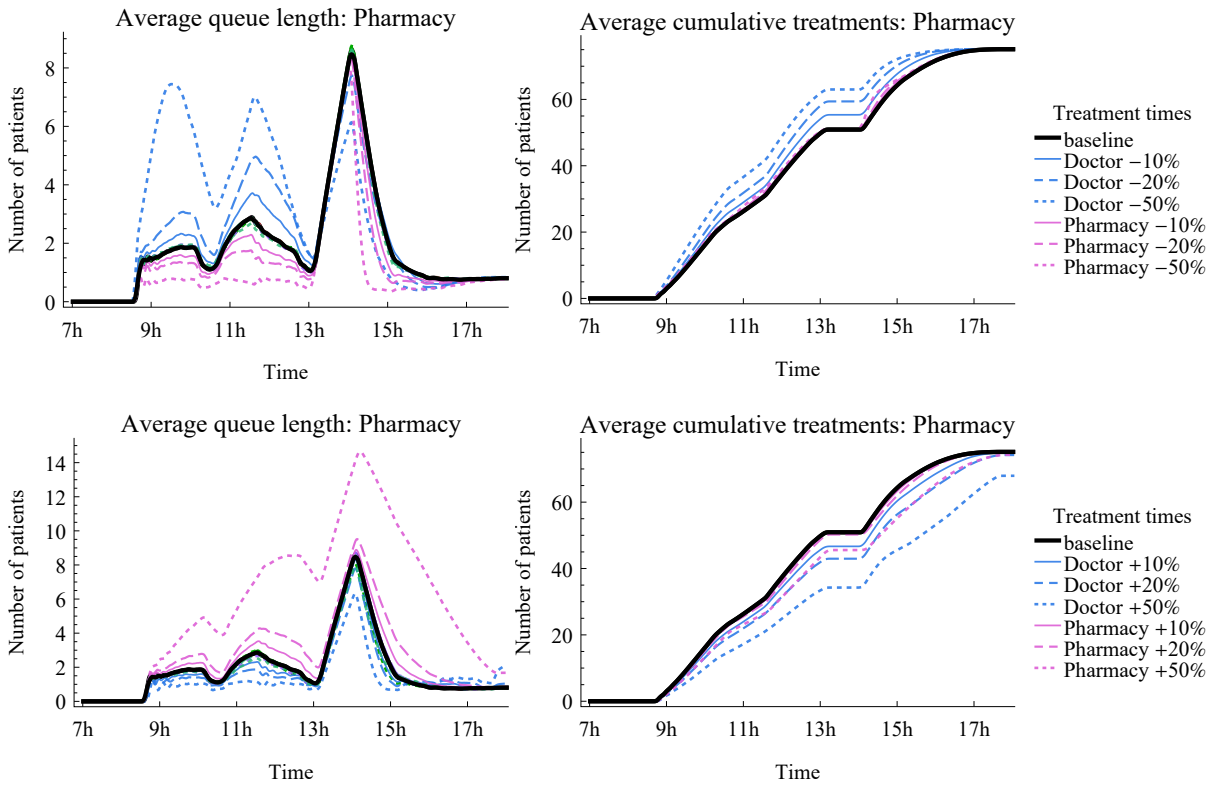


FIGURE 6.16: *The effect of changes to treatment time parameters on the* PHARMACY *queue.*

Stellenbosch University https://scholar.sun.ac.za

6.2 Fluid approximation models

This section contains the results of the two fluid approximation models that were developed in Chapter 3. These results are related to queue length, rather than patient waiting times, since the variables of interest in the fluid models are the expected length of each queue.

Results are provided for both the 2015 and 2016 set-ups, but the purpose of this section is not to evaluate the relative efficiency of the two systems. Instead, the discussion of these results focuses on the following two aspects of the different modelling approaches:

1. How accurately the continuous approximations represent the discrete system (in comparison with the simulation results).
2. Differences between the assumptions of the two fluid models, and how these assumptions are reflected in the results.

6.2.1 Comparison of discrete and continuous models

Figures 6.17 and 6.18 compare the predicted queue lengths from the PF, FCFS and simulation models. As in the previous section, the results of individual simulations are illustrated with thin lines and the average of these results is plotted as a single, thicker line.

The results of the two continuous models are similar to each other, but both models tend to underestimate the queue lengths relative to the simulations. This is particularly apparent over periods where the traffic intensity at a process is lower than 1. In these periods, the continuous models predict a queue length of zero, while the average queue lengths in the simulation results can be as high as three patients.

These discrepancies are due to fluctuations in the queues for individual simulation runs, where the randomness associated with the arrival and treatment times can often lead to several patients arriving at a certain process in quick succession. The continuous models are not able to incorporate these delays because they only consider the average rate of new arrivals.

This problem makes the continuous models less useful than the simulation results when analysing shorter queues or periods of low traffic intensity. For example, the continuous models predict no queues for almost the entire day in the 2015 BLOOD TESTS queue (Figure 6.17) and the 2016 TRIAGE queue (Figure 6.18). The only thing that can be inferred from these results is that the average rate of new arrivals at these processes does not exceed the capacity of the staff on duty, so the overall traffic intensity is less than 1.

By contrast, the simulation results provide more insight into how the behaviour of the queue changes over the course of the day. From these results, it is clear that the traffic intensity at the 2016 TRIAGE queue is highest between about 9h00 and 12h00, which results in an average queue length of at least one patient. Peaks in individual simulation results also indicate that the variance of the queue length is much higher during these times, and queues of more than 5 patients can build up frequently.

The problems with the continuous approximations are most obvious in relatively short queues which have extended periods of low traffic intensity, but they can also have implications for longer queues. The 2015 VITALS queue (Figure 6.17) is an example of how the queue length results in a busy period are skewed by previous periods of low traffic intensity. In this queue, the continuous models correctly identify periods of high traffic intensity in the early morning and during the staff tea break from 10h00 to 11h00. However, the queue length during the latter

Stellenbosch University https://scholar.sun.ac.za

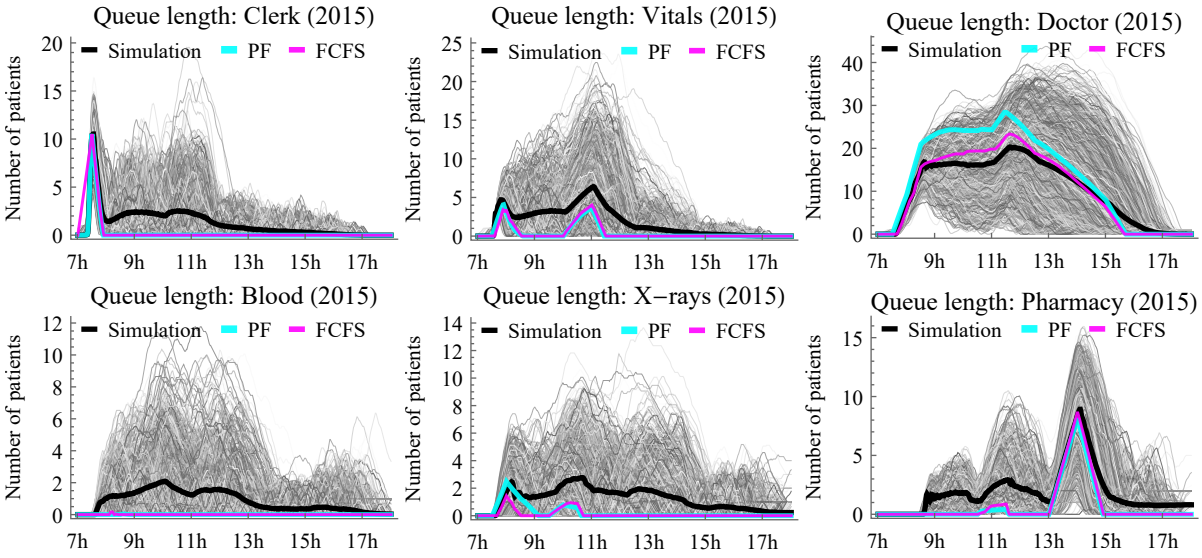


FIGURE 6.17: *A comparison of the average cumulative treatments completed at each process in the 2015 OPD set-up, calculated using the PF, FCFS, and simulation models.*

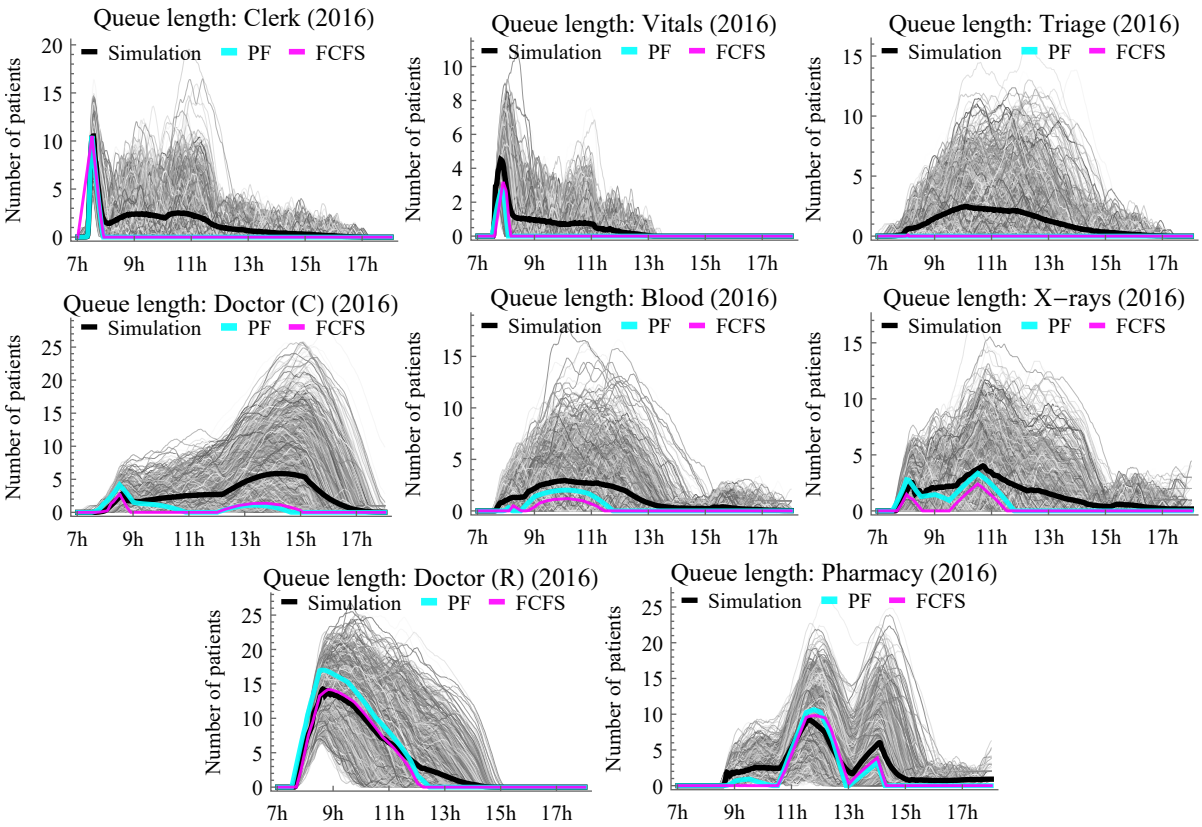


FIGURE 6.18: *A comparison of the average cumulative treatments completed at each process in the 2016 OPD set-up, calculated using the PF, FCFS, and simulation models.*

Stellenbosch University https://scholar.sun.ac.za

interval is dampened by a period of low traffic intensity between 8h30 and 10h00. Although the continuous models predict that the queue will increase between 10h00 and 11h00 and decrease after 11h00, they underestimate the actual length of the queue during this period.

Similar patterns in other processes lead to an interesting observation about the accuracy of the fluid models during periods of high traffic intensity: even when the continuous models provide a poor estimate of the queue length, they are good at predicting whether the queue length is increasing or decreasing.

To illustrate this property, Figure 6.19 shows the rate of change of three OPD queues over the course of the day. The rate of change in queue length is calculated directly from the derivatives of the $q_i^p(t)$ solutions in the fluid models, while the rate of change in the simulation queues is based on numerical derivatives of the average queue lengths over one minute intervals. In these examples, the fluid models identify the same periods of significant increases/decreases in the expected queue lengths as the simulation model.

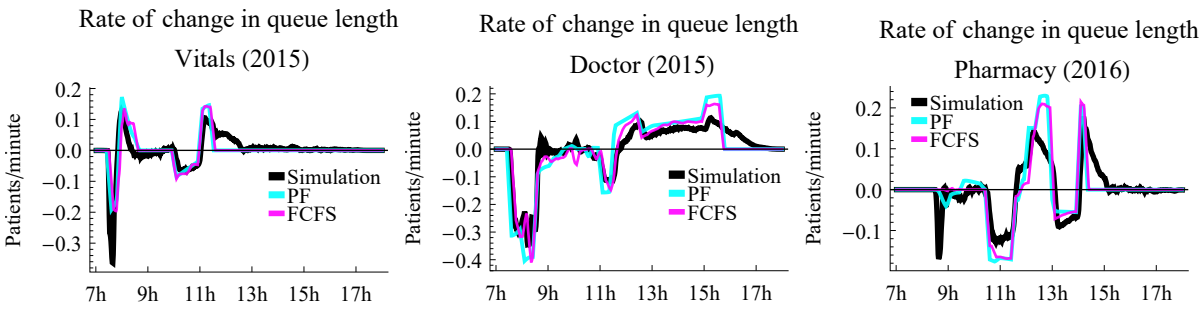


FIGURE 6.19: *Examples of the predicted rate of change in queue length at certain processes in the simulation and fluid models.*

The fluid models' results are also similar to the simulation results in terms of the overall flow of patients through the OPD system. This can be seen in the graphs in Figures 6.20–6.21, which show the expected number of completed treatments that have taken place at each process over the course of the day.

In these graphs, the fluid models give nearly identical results to the simulation model for the CLERKS, VITALS, and TRIAGE processes in both set-ups. Despite the low traffic intensity at these processes, the fluid models are still able to provide an accurate approximation of the flow of patients through these queues. This is a direct contrast to the queue length results for these process, which show significant discrepancies between the discrete and continuous models.

The fluid models tend to overestimate the patient flow slightly at some of the other OPD processes, especially in the 2016 set-up. These differences are linked to the higher variability in patient arrival times at these queues, which causes backlogs and decreases patient flow in the simulation results.

The overall efficiency of the OPD is best measured in terms of the patient flow at the pharmacy. This process is the point of exit for most patients, so the number of patients treated at the PHARMACY over the course of the day is directly related to the amount of time that patients are spending in the OPD. The fluid and simulation models produce very similar results for the 2015 PHARMACY queue, which indicates that these models would predict similar average total waiting times. The results for the 2016 PHARMACY queue are also similar during the morning period, although the fluid models over-estimate the number of patients leaving the OPD between 12h00 and 13h00.

Stellenbosch University https://scholar.sun.ac.za

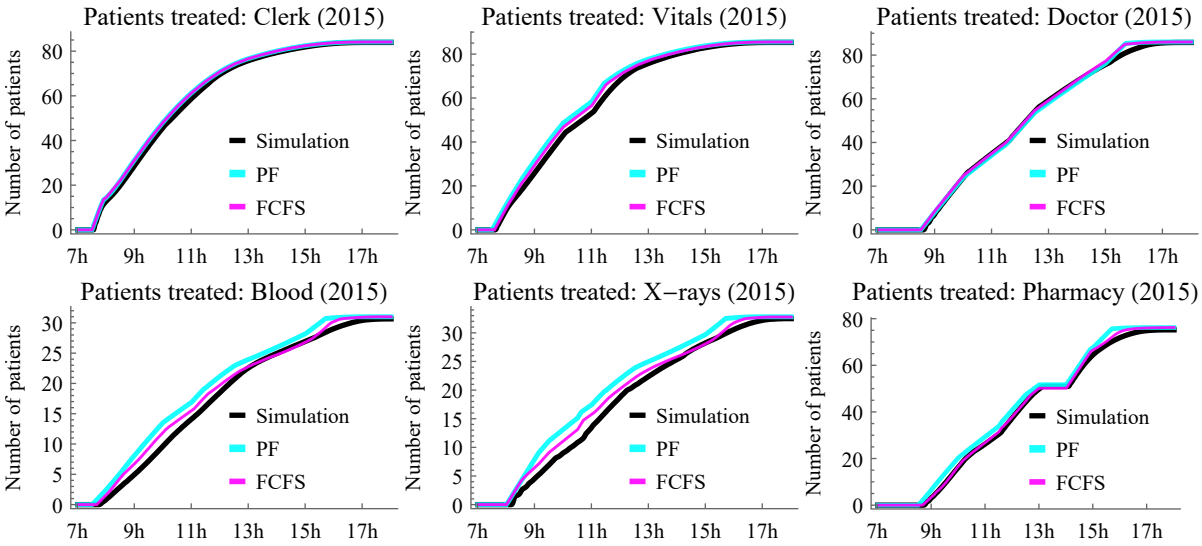


FIGURE 6.20: *A comparison of the expected queue lengths in the 2015 OPD set-up, calculated using the PF, FCFS, and simulation models.*

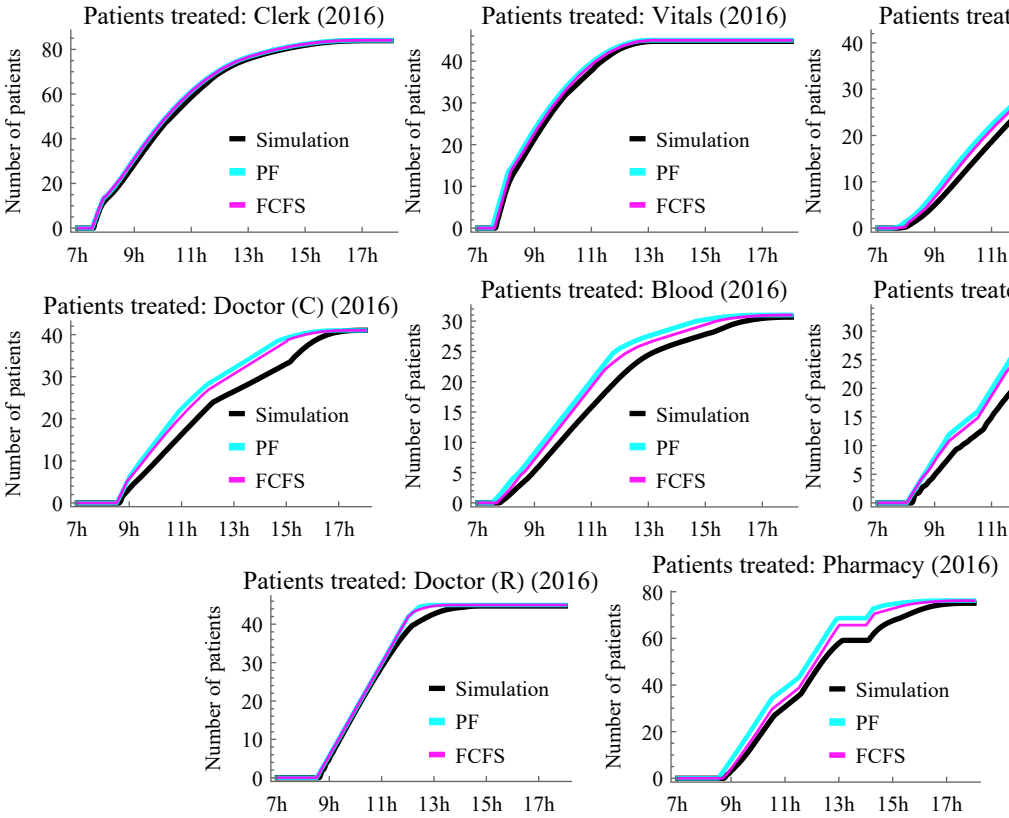


FIGURE 6.21: *A comparison of the expected queue lengths in the 2016 OPD set-up, calculated using the PF, FCFS, and simulation models.*

Stellenbosch University https://scholar.sun.ac.za

Based on these results, the fluid models provide a reasonable approximation of the discrete OPD system. Their results are more limited than the simulation model, since they do not capture the variance in the system or provide estimates of the queue lengths during periods of low traffic intensity. However, they can be used to study the behaviour of very busy queues, to identify periods where the queue lengths are lightly to change rapidly, and to model the flow of patients through the OPD system.

6.2.2 Comparison of fluid models

In this section, the differences in the results of the two fluid models are highlighted. The discussion of these results focusses on intervals of high traffic intensity where these models perform best. Examples from these periods are used to demonstrate how different assumptions in each of the models capture certain aspects of the OPD queues. For convenience, the important differences between the two fluid models are summarised in Table 6.5.

Model	Arrival distribution	Priority queues	FCFS queues	Solution
PF	piecewise	yes	no	numerical
FCFS	continuous	no	yes	analytical

TABLE 6.5: *A summary of the differences between the PF and FCFS models.*

Computational time

Computationally, the PF model is much more efficient than the FCFS model. The PF model results for both OPD set-ups were generated in 3–5 minutes, while the FCFS model results required an hour for the 2016 set-up and several days for the 2015 set-up[1].

The longer computational time for the 2015 set-up is a result of the 2015 routing parameters, which create a loop between the DOCTORS, BLOOD TESTS, and X-RAYS queues. Return patients feed into the DOCTORS queue from the BLOOD TESTS and X-RAYS queues, while casualty patients join the DOCTORS queue first and then proceed to these other processes. This means that the FCFS queues at these processes can only be calculated iteratively in small increments.

Since the FCFS fluid equations are solved analytically through a series of inverted functions, the complexity of these calculations increases significantly with each iteration. Apart from increasing the computational times, these analytical calculations also lead to extremely fragmented piecewise solutions which consist of thousands of individual functions spanning intervals of less than 1 minute. Integrating these functions to find the average queue lengths requires approximately four hours, which makes it very difficult to work with the FCFS model solutions.

Arrival rates

In the PF model, the arrival rates for new patients are modelled by the same triangular distributions that are used in the simulation model. The integrals of these distributions are not suitable for the FCFS model because they contain second-order terms, which are not invertible. Instead of the continuous triangular distributions, the FCFS model uses piecewise-constant versions of these functions that are discretised over 30 minute intervals.

The effect of this discritisation is visible in the results for the early-morning CLERKS queues, which are illustrated in Figure 6.22. The left-hand graph shows that the queue length in the PF model is very similar to the simulation results and correctly predicts the queue growth between

[1]Results were generated on a Dell Optiplex 9010 with Intel I7-3770 CPU (3.4 GHz) and 8GB RAM.

Stellenbosch University https://scholar.sun.ac.za

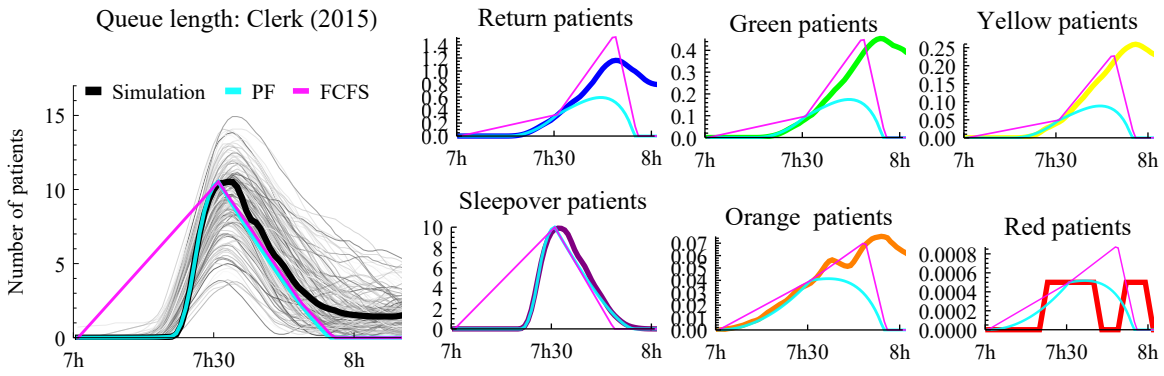


FIGURE 6.22: *Plots of the* CLERKS *queue length between 7h00 and 8h00.*

7h20 and 7h30. The FCFS results are less accurate and predict a linear increase in the queue length from 7h00 to 7h30.

The plots of the individual patient profiles on the right of Figure 6.22 indicate that the discrepancy between the FCFS model and the other results is most severe in the sleepover profile. In the PF and simulation models, sleepover patient arrivals all occur between 7h20 and 7h30, since the arrival distribution for these patients is restricted to this interval. However, the discretised version of this distribution in the FCFS model spreads these arrivals over a 30 minute interval. There is a similar issue with the arrivals of return and green patients, which begin at 7h15 in the PF and simulation models.

One way to address this problem in the FCFS model is to discretise the arrival functions over smaller intervals. Figure 6.23 illustrates the results of the FCFS model with intervals of 30 minutes, 15 minutes, 5 minutes and 1 minute. The results of the 15 minute intervals still deviate from the simulation and PF queues, but both the 5 and 1 minute intervals are accurate. In these cases, the FCFS results for the overall CLERKS queue and the sleepover queue match both the simulation and PF models.

The profile queues for return, green and yellow patients match the PF model, but are slightly different to the simulation results. The smaller arrival intervals have a different effect on the

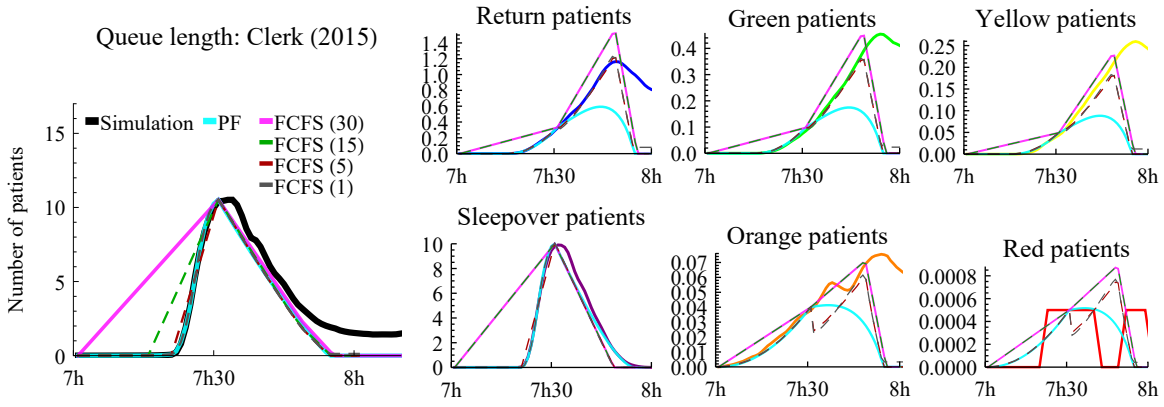


FIGURE 6.23: *The length of the* CLERKS *queue with different arrival function discretisations in the FCFS model.*

Stellenbosch University https://scholar.sun.ac.za

profile queues for orange and red patients. In Figure 6.23, these queues develop a sharp dip just after 7h30 when the interval size is smaller than 15 minutes. These sudden decreases in the queue lengths are related to the queueing disciplines in the fluid models, which are discussed below.

Queue disciplines

Neither of the continuous models fully incorporates the mixture of priority and FCFS queues that are present in the simulation model, but each model is able to reflect certain aspects of this system. The PF model approximates the priority system, while the FCFS model ensures that all patients are processed in the exact order of arrival.

The sharp dip in the profile queue lengths for red and orange patients in Figure 6.23 demonstrates the effect of the queueing discipline in the FCFS model. When the arrival functions are discretised over intervals of 15 minutes or less, red and orange arrivals start at 7h00, while the other profiles only begin arriving after 7h15. The red and orange patients who arrive between 7h00 and 7h15 are ahead of any other profiles in the queue, and are therefore the first patients to be treated when the staff begin work at 7h30. The steep decrease in these queues is the result of these patients leaving the queues just after 7h30.

Since there is no priority in the CLERKS queue, this example does not provide much insight into the effect of the priority discipline in the PF model. To compare the performance of the two models, consider the 2015 DOCTORS queue in Figure 6.24, which has both priority queueing (for red patients) and FCFS queueing (for other profiles).

The continuous models generate similar results for the larger patient profiles in this queue, but the effects of the different queueing disciplines are obvious when it comes to the red and sleepover patient queues in the bottom row of Figure 6.24. For red patients, the PF model gives a smaller average queue length which is much closer to the simulation results than the queue lengths in the FCFS model. In this case, the priority queueing in the PF model is a better reflection of how red patients move through the DOCTORS queue.

The opposite is true for sleepover patients, who tend to get through the DOCTORS queue faster because they arrive earlier than other patients. The FCFS results for the sleepover patients in Figure 6.24 are a very good match for the simulation results, which indicate that all sleepover patients are usually seen by about 10h00. The performance of the PF model is much worse, as it predicts a lingering queue of sleepover patients until about 15h00.

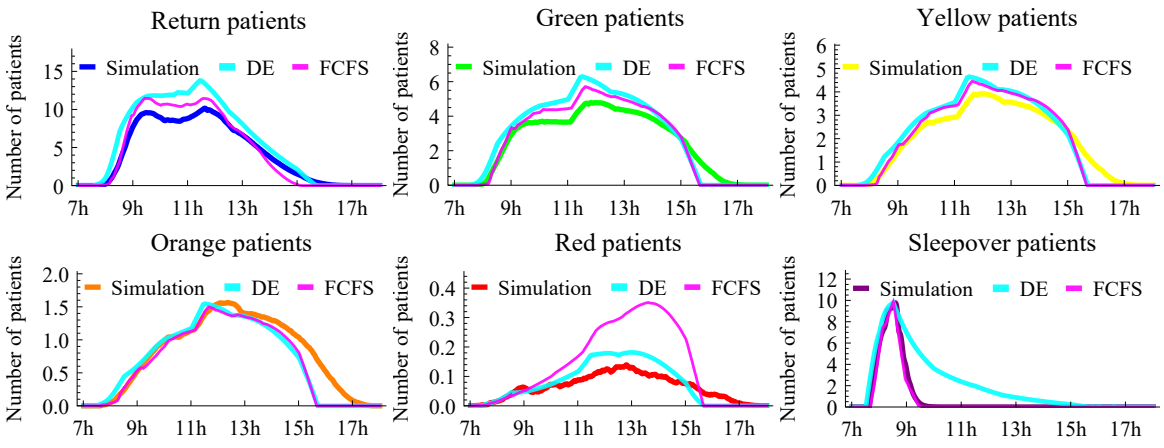


FIGURE 6.24: *A breakdown of the 2015* DOCTORS *queue by profiles. The results for sleepover and red patients highlight the different queueing disciplines in the continuous models.*

Stellenbosch University https://scholar.sun.ac.za

This example is a good illustration of how the queueing disciplines in the continuous models can be both a strength and a weakness, depending on which aspect of the OPD system is being investigated. Each model has specific advantages that are relevant in different circumstances, but neither model is uniformly better or worse than the other. The results of the two models compliment each other — especially when they differ — and are most useful when they are viewed together.

6.3 Summary and recommendations

The results presented in this chapter confirm the concerns expressed by the OPD staff regarding the level of congestion in the facility. Both the simulation and fluid approximation models demonstrate that the DOCTORS queue is a major bottleneck which limits the flow of patients through the system and leads to significant delays. The simulation results also show that the needs of urgent casualty patients require special attention, since these patients experience unacceptably long delays when they are forced to wait in FCFS queues with less urgent patients.

6.3.1 Strategies for addressing the causes of congestion

Based on these results, the root cause of the congestion in the OPD queues is the misalignment of the arrival time distributions and the staff schedules, particularly for the DOCTORS queue. Most OPD patients arrive during the morning, which results in long backlogs of patients that can only be treated when additional doctors are available during the afternoon.

Strategy 1: Redistribute patient arrivals.

In theory, the best way to address this problem would be to set up an appointment scheduling system that could be used to spread the arrivals of return patients more evenly over the day. This would help to avoid overcrowding in the OPD facilities and reduce congestion in the BLOOD TESTS and X-RAYS queues on busy days.

Shifting some of the return patients to the afternoon would also allow doctors to identify and treat urgent casualty patients more efficiently. The current arrival trends are problematic for orange and red patients, because they tend to arrive when the OPD is already busy.

Unfortunately, this strategy is unlikely to be practically implementable. Most of the OPD patients rely on local minibus taxis for transportation to and from the hospital, so they have very little control over their arrival times. It is also likely that patients with afternoon appointments would be hesitant to comply with this system, because they would risk missing the last taxi home and being stranded at the OPD overnight.

Strategy 2: Adjust the OPD staff schedules.

It may be more practical reduce congestion in the OPD by adjusting the staff schedules to match the peak arrival periods. Since the DOCTORS queue plays the most significant role in limiting the flow of patients through the OPD, congestion could be reduced by scheduling additional doctors during the morning and early afternoon.

As demonstrated by the 2016 simulation results, changes to the doctors' schedule will also change the traffic intensity at other processes, so it may also be necessary to adjust these staff schedules. This is not necessarily a straightforward process, as the OPD has limited infrastructure and many of the OPD staff have duties in other parts of the

Stellenbosch University https://scholar.sun.ac.za

hospital. Changes to the staff schedule need to be carefully considered to ensure that they improve patient flow, rather than simply shifting bottlenecks in the network to different processes.

6.3.2 Strategies for addressing the effects of congestion

Even if the OPD can successfully address the causes of congestion, there will still be days when the facility is very busy and some level of congestion is unavoidable. It is therefore important to adopt strategies that mitigate the negative effects of congestion and ensure that patients receive a consistently high standard of care.

Strategy 3: Prioritise urgent patients.

High levels of congestion pose the biggest risk to urgent patients, who are more sensitive to delays. Based on the results in this chapter, introducing priority queues helps to reduce the length of these delays, as well as the amount of variability in the waiting times for urgent patients.

Strategy 4: Avoid overcrowding.

When the OPD is very busy, long queues for particular processes result in overcrowding in certain parts of the OPD. Overcrowding can lead to confusion, and increase the chances of patients standing in the wrong queues. Long queues may also make it difficult for staff members to keep track of individual patients. Based on the simulation results, splitting up very busy processes (like the DOCTORS queue) can help to reduce the maximum queue length on busy days.

The simulation results also indicate that the behaviour of less busy OPD queues can vary significantly from day to day, and even relatively efficient processes can develop long backlogs. In order to minimise the consequences of these backlogs, waiting areas for each process should to be carefully managed and patients need to be given clear instructions about which queues to join.

Strategy 5: Reduce the number of sleepover patients.

Improving patient flow in the OPD can help to reduce the number of sleepover patients. Patients who still need to complete multiple different processes should be prioritised in the morning, especially if they require diagnostic tests that can be completed within one day. It may also help to prioritise patients who live far away from the hospital, as they may have fewer transportation options.

6.3.3 General observations

The results in this chapter highlight a few important factors that need to be taken into account when developing strategies to improve the efficiency of the OPD:

1. The OPD queues function as a network.

   Although it is tempting to focus on the visible signs of congestion — such as long queues — this is not necessarily the best way to reduce patient waiting times. Improving the efficiency of a particular process can result in longer queues elsewhere, so it is important to consider each process in terms of the overall flow of patients through the system.

Stellenbosch University https://scholar.sun.ac.za

2. Congestion does not affect all patient profiles in the same way.

   The arrival distributions, treatment needs, and routing of different patient profiles can have a significant effect on their experiences in the OPD. Balancing the needs of the different patient profiles is important, since changes that benefit one profile can disadvantage many others.

3. More data is needed.

   The sensitivity analysis in § 6.1.5 indicates that the treatment times at the DOCTORS queue play a very significant role in determining the flow of patients through the OPD system. It is therefore important to have an accurate understanding of these treatment times, as well as other factors that influence the efficiency of this queue.

   On a more general level, the OPD also needs to develop data collection strategies to monitor the efficiency of the OPD on an ongoing basis. This data should focus on patient waiting times, since visual assessments of the OPD queues are not necessarily a good measure of patient flow.

Stellenbosch University https://scholar.sun.ac.za

CHAPTER 7

Optimisation

This chapter introduces an optimisation model for the staff schedules in the OPD network. The optimisation model is based on the conceptual model in Chapter 2, and focusses on improving the flow of patients through the network. The variables, constraints and assumptions of the optimisation model are discussed in § 7.1–§ 7.2.

The optimisation is implemented as a genetic algorithm, which is combined with the OPD simulation model. Section 7.3 describes how the simulation model is used to evaluate the efficiency of different OPD set-ups and § 7.4 describes the genetic algorithm procedures for generating improved staff schedules.

Section 7.5 contains an overview of the tests that were conducted to determine the appropriate parameters for the genetic algorithm. A detailed discussion of these results is provided to demonstrate how different combinations of parameters affect both the efficiency of the algorithm and the quality of the final solutions. Three of the best solutions found by the genetic algorithm are compared to the 2015 OPD staff schedule in § 7.6.

7.1 Variables

The scope of the optimisation model is limited to variables that are within the control of the hospital, so many of the factors that contribute to congestion in the queueing network are considered to be fixed. This includes all of the information and parameters contained in the patient profiles: the number of patients, their arrival patterns and their treatment needs. The OPD processes and the structure of the queueing network are also kept constant, since these factors are linked to the hospital's infrastructure and determined by high-level operational policies that cannot be changed.

The focus of the optimisation is the hospital's staff schedules, which determine the distribution of staff across different processes during the day. These schedules are generated on a weekly basis, so they can be changed by the OPD staff. If the total number of staff-hours is kept constant, this can also be done at no cost to the hospital.

The staff schedule is divided into T equal time intervals which span the OPD working hours. The number of staff on duty at any process i is given by the variable $x_{i,t}$, where t indicates the time interval. A complete schedule consists of a set of $n \times T$ variables,

$$\mathbf{x} = (x_{1,1}, \ldots, x_{1,T}, \ldots, x_{i,1}, \ldots, x_{i,T}, \ldots, x_{n,1}, \ldots, x_{n,T}). \tag{7.1}$$

Stellenbosch University https://scholar.sun.ac.za

7.2 Constraints and assumptions

The optimisation model assumes that the maximum number of patients that can be treated concurrently at each process is limited by the OPD infrastructure. This includes spatial restrictions such as the number of desks or consultation rooms allocated to each process, as well as the availability of specialist equipment needed for certain processes. Increasing the number of staff beyond these limits does not increase the number of patients that can be treated, so the number of staff assigned to a process at any given time in the day is limited by the constraint

$$x_{i,t} \leq b_i, \qquad \text{with } i \in \mathcal{I} \text{ and } t \in \{1, \ldots, T\}, \tag{7.2}$$

where b_i represents the maximum number of patients that can be accommodated by the infrastructure allocated to process i.

There are many different types of staff who work in the OPD, such as doctors, nurses, other medical professionals, and various administrative staff. To avoid assigning the wrong type of staff member to a specific process, it is assumed that each process has its own independent set of s_i staff. This leads to the second constraint,

$$x_{i,t} \leq s_i, \qquad \text{with } i \in \mathcal{I} \text{ and } t \in \{1, \ldots, T\}, \tag{7.3}$$

which restricts the maximum number of staff available for each process.

It is also assumed that the total number of staff-hours worked at each process should not be increased, since this would require additional funding to compensate staff who work longer hours. The maximum number of hours worked at each process is limited by the constraint

$$\sum_{t=1}^{T} x_{i,t} \leq h_i, \qquad \text{with } i \in \mathcal{I}. \tag{7.4}$$

If certain staff members can work at multiple processes, the OPD processes are grouped into subsets $\mathcal{J}_1, \mathcal{J}_2, \ldots, \mathcal{J}_K$ according to the different staff types. The staff constraints for these processes can be reformulated as

$$\sum_{j \in \mathcal{J}_k} x_{j,t} \leq S_k, \qquad \text{with } t \in \{1, \ldots, T\} \text{ and } \; k \in \{1, \ldots, K\}, \tag{7.5}$$

$$\sum_{j \in \mathcal{J}_k} \sum_{t=1}^{T} x_{j,t} \leq H_k, \qquad \text{with } k \in \{1, \ldots, K\}, \tag{7.6}$$

where S_k is the total number of staff members that are shared between the processes in $\mathcal{J}_k$ and H_k is the maximum amount of time that these staff members may work.

Additional constraints of the form $x_{i,t} \leq u_{i,t}$ may be added to this model to reflect changes in the availability of staff members at specific times of the day, while constraints in the form $x_{i,t} \geq l_{i,t}$ can be used to ensure that the number of staff on duty does not drop below a certain level. In some cases, it may also be useful to include constraints that limit the number of staff changes between successive time intervals, i.e.

$$\sum_{t=1}^{T-1} \|x_{i,t+1} - x_{i,t}\| \leq c_i, \qquad i \in \mathcal{I}. \tag{7.7}$$

Stellenbosch University https://scholar.sun.ac.za

7.3 Objective function

The efficiency of different staff schedules is compared using the results of the OPD simulation model. There are many different ways to measure the efficiency of each OPD set-up, so the objective function for the optimisation algorithm could include various different statistics from the simulation results.

Most of these statistics are related to waiting times for the different patient profiles, so the variables $w_{p,k}$ are used to represent the total waiting time for the k^{th} patient from profile p. The three main waiting time statistics that will be considered are

$$z_1 = \frac{1}{\eta} \sum_{p=1}^{m} \sum_{k=1}^{\eta_j} w_{p,k}, \tag{7.8}$$

$$z_2 = \frac{1}{\eta} \sum_{p=1}^{m} \sum_{k=1}^{\eta_j} \theta\Big(w_{p,k} - w_p^{(u)}\Big), \text{ and} \tag{7.9}$$

$$z_3 = \frac{1}{z_2} \sum_{p=1}^{m} \sum_{k=1}^{\eta_j} \theta\Big(w_{p,k} - w_p^{(u)}\Big) \frac{w_{p,k}}{w_p^{(u)}}. \tag{7.10}$$

Equation (7.8) gives the average total waiting time for all OPD patients. Minimising the average waiting time per patient is a simple way to improve the efficiency of the OPD, and a decrease in this figure would be beneficial to the majority of patients. However, it does not take into account that the consequences of an average length wait are likely to differ significantly between different types of patients.

This problem is addressed by considering the waiting time targets for different patients. In equation (7.9), the function $\theta(\cdot)$ is the Heaviside step function which returns a value of 1 for positive arguments and 0 otherwise. The statistic z_2 gives the total number of OPD patients not treated within the appropriate target times, which incorporates the fact that certain patients are more urgent than others. However, minimising this statistic is likely to result in solutions that favour the larger patient profiles, since their contribution to this statistic is much higher.

This is not the case for equation (7.10), which calculates the average factor by which these waiting time targets are exceeded. This gives more weight to the most urgent profiles, since they have the shortest target times. Minimising z_3 on its own will not necessarily improve the efficiency of the system, since schedules that result in a large number of targets being missed by a very small margin would be preferred over schedules that only miss one or two targets by a slightly higher margin.

Other statistics that may be useful in the objective function are the maximum waiting time or maximum deviation from the target times, as well as the maximum number of missed targets on any given day. Minimising these statistics may increase the fairness of the schedule by ensuring that individual patients are not severely disadvantaged in order to improve the overall performance.

Minimising the variance of these statistics can also help to produce schedules that give a relatively consistent level of service, despite the daily fluctuations in patient arrivals and treatment times. A solution that produces moderate results on both busy and quiet days is preferable to a solution that gives excellent results on quiet days and very poor results on busy days.

Due to the inherent variability in the simulation results, the objective function is unlikely to provide an accurate comparison between two potential solutions unless a large number of sim-

Stellenbosch University https://scholar.sun.ac.za

ulations are run. However, increasing the number of simulations also increases the amount of time needed to calculate the objective function and reduces the number of potential solutions that can be compared.

One way to limit the variability of the simulation results and increase the chances of a fair and accurate comparison is to use exactly the same simulations to test each solution. To do this, patient arrivals and treatment times are randomly generated for S different simulations. Each potential solution is tested on all of the S simulations by implementing the different schedules without changing any of the patient samples. Since the schedule is the only variable that changes between different runs of these simulations, their results are much more likely to give valid comparisons of different schedule solutions.

This method may have certain disadvantages, since there is a danger that the final solution will be optimised to fit the set of S simulations, instead of the underlying parameters in the OPD set-up. This can be problematic if the simulations used for the optimisation are a poor representation of the normal behaviour of the system, as the final solution may perform poorly in more typical circumstances. It is therefore important to ensure that S is large enough to provide a representative sample. To achieve this, the size of S must be greater when there are high levels of variability in either the patient data or the waiting time statistics in the objective function.

7.4 Genetic algorithm

The schedule optimisation is performed using a genetic algorithm (GA), a meta-heuristic optimisation technique that mimics the process of natural selection. The algorithm begins with an initial generation of potential solutions, which are randomly generated, and then iteratively improves the solutions in successive generations until the stopping criteria is reached.

The initial generation consists of the current staff schedule as well as $N-1$ random alternative solutions. Alternative solutions are generated by starting with an empty schedule, $\mathbf{x} = (0, \ldots, 0)$, and randomly distributing the total staff time for each process, h_i, across the variables $x_{i,t}$. This is done by randomly choosing variables from the set $\{x_{i,t} \| x_{i,t} < \min[s_i, b_i, u_{i,t}]\}$ and increasing these variables by one unit until the total number of staff hours have been assigned.

The solutions in the initial generation are ranked in terms of their *fitness values*, which are calculated using the objective function. The algorithm then discards the weakest solutions and replaces them with new solutions to form a new generation. These new solutions are called *children*, because they are generated by combining pairs of *parent* solutions which are selected from the fittest solutions in the current generation.

The aim of the genetic algorithm is to mimic the process of natural selection by retaining the characteristics of fitter solutions and discarding weaker solutions in each new generation. This procedure consists of three steps: *cross-over, mutation* and *immigration*. A detailed explanation of each of these steps is provided below in § 7.4.1-7.4.3 and the stopping criteria for the algorithm is discussed in § 7.4.4

7.4.1 Cross-over

In the cross-over step, $M/2$ pairs of parent solutions are selected from the current generation and combined to produce M new children. Each pair of parents contains one of the strongest solutions of the current generation, so the fittest $M/2$ solutions are automatically selected for

Stellenbosch University https://scholar.sun.ac.za

crossover. The second parent in each pair is selected from among the remaining solutions in the current generation.

The normal procedure for generating child solutions in the genetic algorithm is to randomly select a set of variables in one parent solution and replace these variables with the corresponding values from the other parent solution. In this case, however, the cross-over operation has to be performed more carefully, since switching a random set of variables from two parent solutions is very likely to produce a child that breaks the constraint on the maximum number of staff-hours.

To avoid this problem, the number of staff-hours for each process is fixed by the upper bounds in the optimisation constraints, so that

$$\sum_{t=1}^{T} x_{i,t} = h_i, \qquad \text{with } i \in \mathcal{I}. \tag{7.11}$$

The cumulative sum of each solution vector is calculated in a new vector of $n \times T$ variables,

$$\mathbf{y} = (y_1, \ldots, y_T, \ldots, y_{2\times T}, \ldots, y_{i\times T}, \ldots, y_{n\times T}), \qquad \text{where } y_k = \sum_{j=1}^{k} x_j. \tag{7.12}$$

If each process has been allocated exactly h_i staff hours, then the staff hour constraints can be re-written as

$$y_{i\times T} = \sum_{j=1}^{i} h_j, \qquad \text{with } i \in \mathcal{I}. \tag{7.13}$$

For each pair of parent solutions, $\mathbf{x}^{(1)}$ and $\mathbf{x}^{(2)}$, a difference vector is calculated by subtracting the cumulative sum vectors of the two solutions, i.e.

$$\mathbf{d} = \mathbf{y}^{(1)} - \mathbf{y}^{(2)} = (d_1, \ldots, d_{i\times T}, \ldots, d_{n\times T}). \tag{7.14}$$

Sequences of consecutive values $(x_a, x_{a+1}, \ldots, x_b)$ can be transplanted from $\mathbf{x}^{(1)}$ into $\mathbf{x}^{(2)}$ provided that $d_b = 0 = d_{a-1}$, since the total number of staff hours in these sequences is the same in both parent solutions. Switching sequences of variables that fulfil this condition will not change the overall number of staff hours assigned to each process in either parent solution.

There will always be at least n different sections in each pair of solutions that meet this requirement, since

$$y^{(1)}_{k\times T} = y^{(2)}_{k\times T} = \sum_{i=1}^{k} h_i. \tag{7.15}$$

There are also likely to be many more points of intersection that occur within these sections, resulting in small sequences of two or three values. To avoid switching two identical sequences of variables, sequences are only switched if there is at least one $d_c \neq 0$ for $a \leq c < b$.

To illustrate the cross-over process, consider the following two solutions:

$$\mathbf{x}^{(1)} = (2,2,2,2,3,3,3,3,3,1,1,1,2,2,2) \quad \mathbf{y}^{(1)} = (2,4,6,8,11,14,17,20,23,24,25,26,28,30,32)$$
$$\mathbf{x}^{(2)} = (1,2,2,3,4,2,2,2,2,2,3,3,2,2,0) \quad \mathbf{y}^{(2)} = (1,3,5,8,12,14,16,18,20,22,25,28,30,32,32)$$

The match vector for this pair of solutions indicates that the solutions intersect at four points:

Stellenbosch University https://scholar.sun.ac.za

$$\mathbf{d} = \mathbf{y}^{(1)} - \mathbf{y}^{(2)} = (\underbrace{1,1,1,0,}\ \overbrace{-1,0,}\ \underbrace{1,2,3,2,0,}\ \overbrace{-2,-2,-2,0})$$

$$\mathbf{x}^{(1)} = (\underbrace{2,2,2,2,}\ \overbrace{3,3}\ ,\underbrace{3,3,3,1,1,}\ \overbrace{1,2,2,2}) \qquad \mathbf{x}^{(2)} = (\underbrace{1,2,2,3,}\ \overbrace{4,2}\ ,\underbrace{2,2,2,2,3,}\ \overbrace{3,2,2,0})$$

To produce two new children, every second section in each of the parent solutions is replaced by the corresponding section from the other parent.

$$\mathbf{c}^{(1)} = (\underbrace{2,2,2,2,}\ \overbrace{4,2}\ ,\underbrace{3,3,3,1,1,}\ \overbrace{3,2,2,0}) \qquad \mathbf{c}^{(2)} = (\underbrace{1,2,2,3,}\ \overbrace{3,3}\ ,\underbrace{2,2,2,2,3,}\ \overbrace{1,2,2,2})$$

It may occur that two parent solutions are very similar and cannot produce more than one distinct sequence of switching variables, so the child solutions would be the same as the two original parents solutions. In this case, a different solution is substituted for the weaker parent until a better match is found.

The major advantage of this cross-over method is that it will always produce feasible children, but it is also beneficial because it can preserve important links between the schedules for different processes in the parent solutions. Strong solutions tend to encourage a steady flow of patients through the system, so the number of staff assigned to each process at a given point in the day is often aligned with the schedules at the previous process. By alternating sections from two parents, this cross-over method avoids significant disruptions to this flow.

7.4.2 Mutation

Once a successful cross-over has been performed, small alterations are made to a certain proportion of the child solutions. The probability of a particular solution being mutated is given by the parameter u. Mutations are performed by choosing a random, non-zero variable in the solution and decreasing the value of that variable by 1. Another variable is then selected from the same process and increased by 1 to balance the schedule, for example

$$\mathbf{c}^{(1)} = (2,2,2,\underline{2},4,3,3,3,1,1,\underline{2},2,2,0) \quad \longrightarrow \quad \mathbf{c}^{(1^*)} = (2,2,2,\underline{3},4,3,3,3,1,1,\underline{1},2,2,0)$$

If the second variable is already set to the maximum number of staff, new variables are selected until a smaller value is found.

7.4.3 Immigration

After all cross-over and mutation operations have been completed, the final step in this process is to identify any children which are an exact match to other solutions in the current generation. These solutions are discarded and replaced with randomly generated solutions to maintain adequate diversity in the population. After calculating the fitness values for all child solutions, a new generation is formed by replacing the parent solutions in the current generation with the child solutions.

Stellenbosch University https://scholar.sun.ac.za

7.4.4 Stopping criteria

The best solution found in each iteration of the genetic algorithm is stored, and its fitness value is compared to the best solutions from all previous generations. The algorithm is terminated when there is no improvement in the best solution for ten consecutive generations.

7.5 Parameters

The parameters in the genetic algorithm must be carefully selected to ensure that the algorithm is both effective (produces good results) and efficient (runs quickly). A series of tests were run using the 2015 OPD set-up to determine the appropriate combination of parameters for this particular problem. The following parameter values were tested:

1. Population size: $N \in \{10, 20, 30, 40, 50\}$
2. Number of children: $M \in \{0.2N, 0.35N, 0.5N, 0.65N\}$
3. Mutation probability: $u \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$
4. Number of simulations: $S \in \{5, 10, 15, 20, 25, 30\}$

In addition to these parameters, three different objective functions were also considered:

1. Average waiting time:

$$f_1 = a_1 \frac{1}{N} \sum_{p=1}^{m} \sum_{k=1}^{\eta_p} w_{p,k}. \tag{7.16}$$

2. Average number of targets missed:

$$f_2 = a_2 \sum_{p=1}^{m} \sum_{k=1}^{\eta_p} \theta\left(w_{p,k} - w_p^{(u)}\right). \tag{7.17}$$

3. Average number of targets missed + maximum deviation from targets:

$$f_3 = \frac{1}{2}\left(f_2 + \frac{a_3}{z_2} \sum_{p=1}^{m} \sum_{k=1}^{\eta_j} \theta\left(w_{p,k} - w_p^{(u)}\right) \frac{w_{p,k}}{w_p^{(u)}}\right). \tag{7.18}$$

The coefficients a_1, a_2 and a_3 were calculated based on the fitness value for the initial solution, $\mathbf{x}_0$. These coefficients scale the initial fitness value to 100, i.e.

$$f_1(\mathbf{x}_0) = 100, \;\; f_2(\mathbf{x}_0) = 100, \quad \text{and } f_3(\mathbf{x}_0) = 100. \tag{7.19}$$

Since each run of the genetic algorithm generates a new set of S simulations for the objective function, the values of a_1, a_2 and a_3 were recalculated for each run. These coefficients do not affect the ranking of solutions within a single run, since every fitness value in a particular run is multiplied by the same coefficient. However, the coefficient values are relevant when comparing the fitness values from multiple different runs of the algorithm which were generated using different simulations.

The values of s_i and b_i for each process were set to the maximum number of staff at that process in the 2015 set-up, and the number of staff-hours scheduled for each process was kept constant. No constraints were included for tea breaks or lunch breaks, but working hours for all staff were

Stellenbosch University https://scholar.sun.ac.za

restricted to the period 7h00–18h30. Additional constraints were added for doctors, who attend a staff meeting until 8h30.

7.5.1 Computational time and efficiency

In the genetic algorithm, the simulation runs are far more time-consuming than other operations in the rest of the algorithm. The number of simulations used in the fitness function has a strong influence on the computational time, which increases linearly with S. The total number of simulation runs performed in a single run of the genetic algorithm is $S \times (N + GM)$, where G is the number of generations before the stopping criteria are reached. The average computational times for different combinations of N, M and S are shown in Table 7.1.

M=0.2N	N=10	N=20	N=30	N=40	N=50
S=5	13	55	85	121	159
S=10	27	116	193	267	346
S=15	46	169	293	401	520
S=20	59	223	406	544	696
S=25	69	271	489	654	840
S=30	75	346	595	793	1060

M=0.35N	N=10	N=20	N=30	N=40	N=50
S=5	17	68	128	413	243
S=10	31	146	284	423	531
S=15	45	213	442	632	814
S=20	64	307	574	849	1050
S=25	86	357	754	1052	1380
S=30	84	430	892	1241	1619

M=0.5N	N=10	N=20	N=30	N=40	N=50
S=5	15	44	100	136	180
S=10	34	101	220	292	406
S=15	46	150	335	437	627
S=20	58	205	449	578	863
S=25	73	239	555	732	1079
S=30	93	267	668	852	1261

M=0.65N	N=10	N=20	N=30	N=40	N=50
S=5	7	17	38	81	115
S=10	16	33	92	169	261
S=15	21	50	141	263	397
S=20	29	62	200	362	521
S=25	35	86	244	431	660
S=30	42	109	252	530	784

TABLE 7.1: *The average computational times (in seconds) for the genetic algorithm using different parameters.*

In addition to the computational time, the efficiency of the algorithm also depends on the diversity of the population in each iteration. The population diversity is measured in terms of the average percentage of new solutions in each iteration that are an exact match for one of the existing solutions in the current population. The algorithm progresses more slowly when there are a large number of repeated solutions, since these solutions must replaced with randomly generated solutions that lower the overall quality of each generation.

The charts in Figure 7.1 show the average number of repeated solutions in each generation for the different combinations of parameters that were tested. In these results, the mutation probability u has the strongest influence on the number of repeated solutions, and the most effective way to avoid this problem is to use a very high mutation probability, $u \geq 0.7$.

The parameters N and M also influenced the average number of repeated solutions in each generation, although their effects were not as dramatic as the mutation parameter. Repeated solutions occurred less frequently with large populations, $N \geq 30$, and the smallest population size ($N = 10$) also performed well, especially when $u \leq 0.5$. Populations of 20 and 30 solutions had the highest number of repeated solutions.

In all cases, the smallest number of children $M = 0.2N$ resulted in significantly more repeated solutions than the other cases that were tested, so the number of children in each generation

Stellenbosch University https://scholar.sun.ac.za

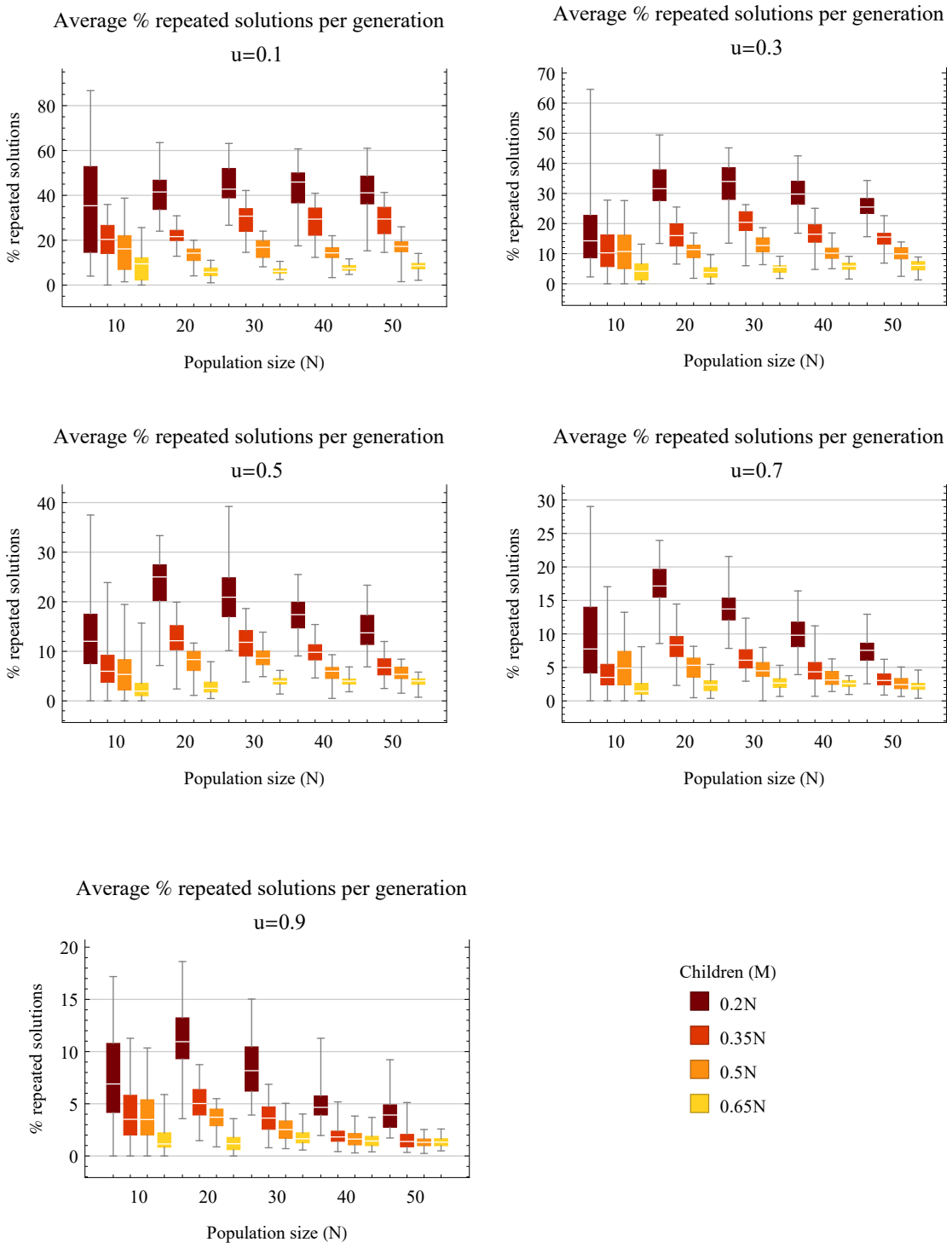


FIGURE 7.1: *A comparison of the average number of repeated solutions in each iteration of the genetic algorithm.*

Stellenbosch University https://scholar.sun.ac.za

should be at least 30–40% of the total population size. The number of repeated solutions decreased for higher values of M, but it is not practical to set this parameter too high because it can become difficult to find the required number of feasible parent pairs in each generation when $M > 0.6N$.

7.5.2 Solution quality

The number of simulation runs used to evaluate the fitness function influences both the efficiency and the effectiveness of the algorithm. If too few simulation runs are used, the fitness function will be inaccurate and the algorithm is more likely to produce poor quality solutions. However, using a large number of simulation runs reduces the efficiency of the algorithm and may require a decrease in the population size.

In the tests that were conducted using different parameter values, the accuracy of the fitness functions was tested by re-evaluating the solutions in the last generation of each run of the genetic algorithm. The fitness values for these solutions were recalculated using a new, independent sets of simulations.

These new fitness values were then compared to the initial fitness values using Spearman's rank correlation coefficient (Spearman, 1904). A summary of these coefficients is illustrated in Figure 7.2. A value of $\rho = 1$ indicates that both sets of fitness values rank each of the solutions in the same order, while $\rho = 0$ indicates that there is no consistency between the original fitness values and the new fitness values.

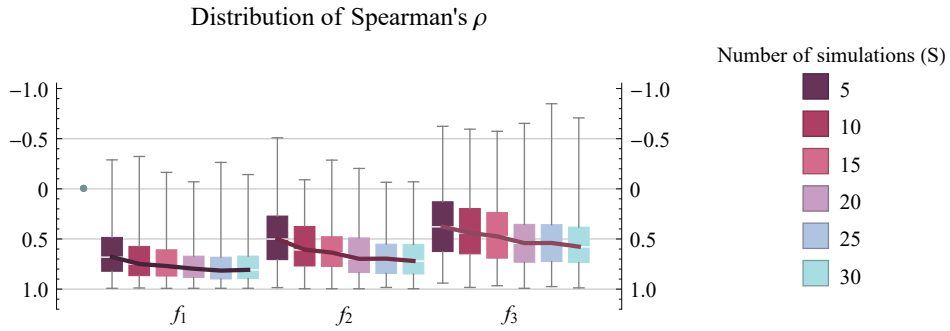


FIGURE 7.2: *Spearman's rank correlation coefficient for the fitness function values generated using different sets of simulations.*

All three fitness functions performed best with the highest number of simulations that was tested, $S = 30$, but these results were not significantly better than cases where $S = 25$. In all cases, the first objective function ranked different solutions more consistently than the other two functions, and the third objective function performed worst.

The effect of the parameters N, M, and u on the performance of the genetic algorithm was investigated using the best solutions found in test cases where $S = 25$ or $S = 30$. The fitness values for these solutions were re-calculated using a set of 100 simulations to ensure an accurate comparison. A summary of these fitness values for different parameters is shown in Figure 7.3.

For both the first and second objective functions, the parameters $u \geq 0.7$, $N \geq 30$ and $M \leq 0.5N$ produced the best solutions. The range of fitness values observed for these parameters in the test cases was very small, which indicates that the performance of the algorithm was consistently good. For populations smaller than 30, the lack of diversity in the population caused the algorithm to converge very quickly, which resulted in poorer final solutions. The algorithm was

Stellenbosch University https://scholar.sun.ac.za

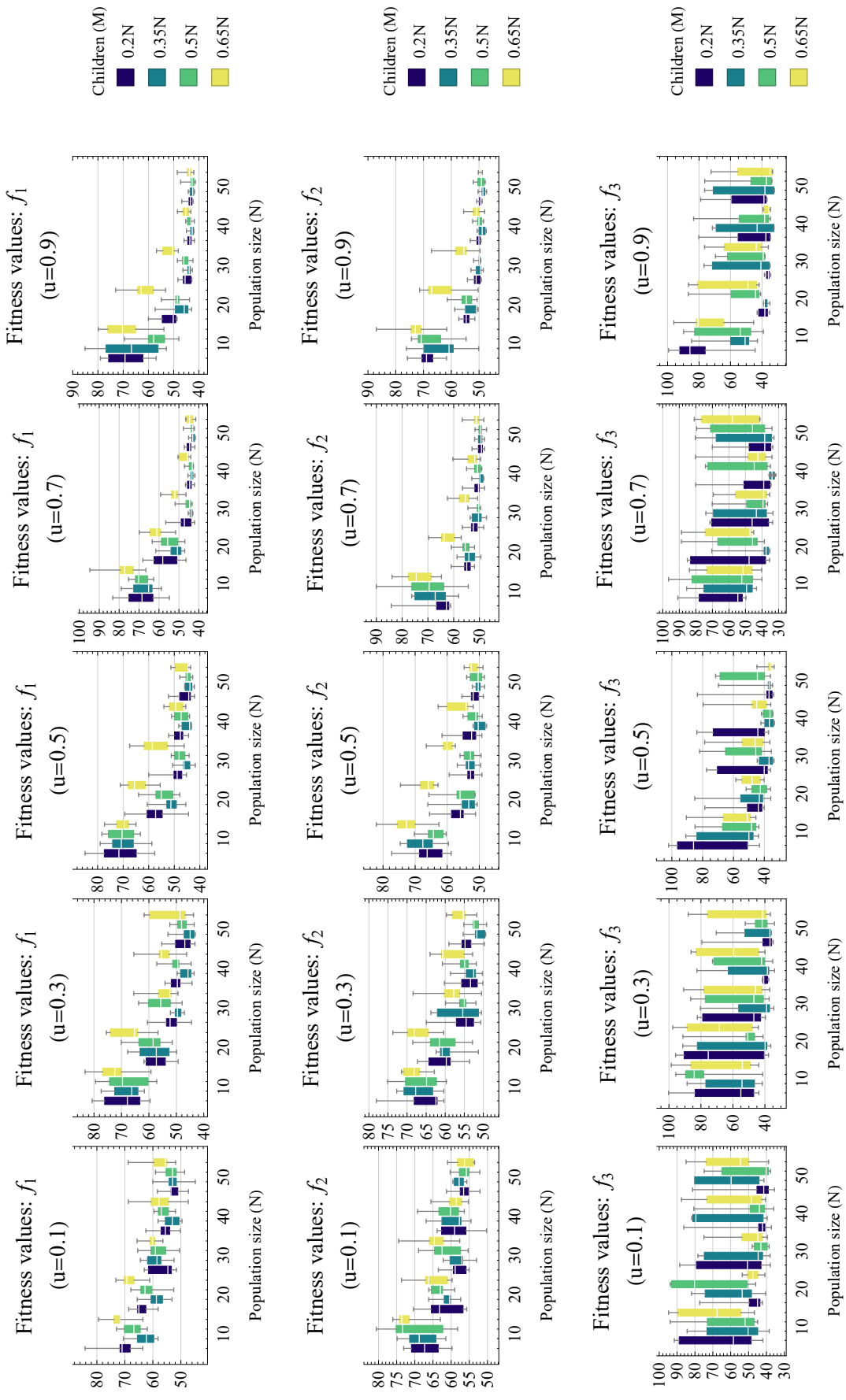


FIGURE 7.3: *A comparison of the objective function values achieved using different numbers of children, population sizes and mutation probabilities.*

Stellenbosch University https://scholar.sun.ac.za

also less effective when the number of children was large ($M = 0.65N$) as it was more difficult to find feasible pairs of parent solutions.

The performance of the genetic algorithm was very erratic for the third objective function due to the high levels of variance in the maximum deviations from waiting time targets. Figure 7.4 shows that the best solutions found using the third objective function did perform relatively well in terms of the average deviations from the waiting time targets, although they had higher average waiting times and a higher number of missed targets than the solutions found by the other two objective functions.

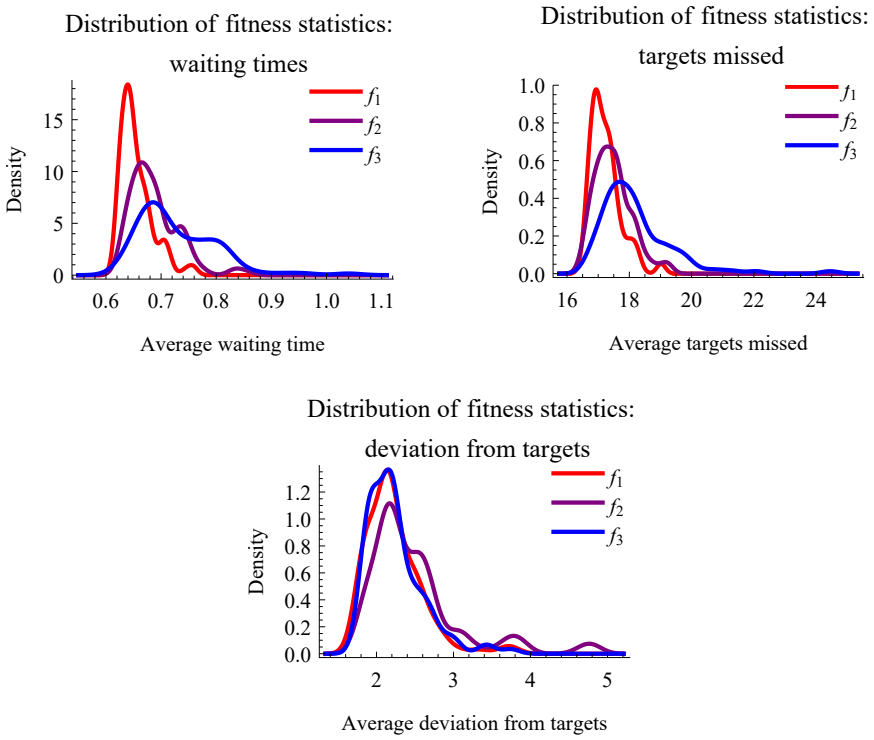


FIGURE 7.4: *A comparison of the distribution of waiting time statistics for the best OPD staff schedules found by the genetic algorithm.*

As expected, Figure 7.4 shows that the first objective function produced the best average waiting times. It also outperforms the second objective function in terms of the number of targets missed, and produced very similar results to the third objective function for the average deviation from targets. Based on these results, the average waiting times in the OPD were the best measure of the efficiency of different staff schedules in the genetic algorithm.

7.6 Results

In this section, the best solutions found by each of the three objective functions are compared to the 2015 OPD set-up. The staff schedules for these solutions are shown in Table 7.2. In all three of these schedules, the number of staff working during the morning has increased relative to the original schedule and there are fewer staff scheduled to work during the afternoon. The new schedules were tested using the first 1000 simulations that were generated to compare the 2015 and 2016 set-ups in Chapter 6.

Plots of the average queue length at different processes (Figure 7.5) show that the three new

Objective function 1

Process	7h30	8h	8h30	9h	9h30	10h	10h30	11h	11h30	12h	12h30	13h	13h30	14h	14h30	15h	15h30	16h	16h30	17h	17h30	18h
CLERKS	3	2	3	3	3	3	3	3	3	2	2	2	3	3	2	3	3	2	3	1	0	1
VITALS	0	0	2	2	2	2	2	2	2	2	2	2	2	2	1	1	1	0	1	2	0	0
DOCTORS	0	0	4	4	4	4	4	4	4	4	3	2	1	3	4	1	0	1	2	1	0	1
BLOOD TESTS	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
X-RAYS	0	2	2	2	2	2	2	2	2	2	2	2	1	1	1	1	1	1	2	1	0	0
PHARMACY	2	2	2	2	2	2	2	2	2	2	2	1	2	1	2	1	0	1	2	1	1	2

Objective function 2

Process	7h30	8h	8h30	9h	9h30	10h	10h30	11h	11h30	12h	12h30	13h	13h30	14h	14h30	15h	15h30	16h	16h30	17h	17h30	18h
CLERKS	3	2	3	2	3	3	2	1	3	3	3	3	2	3	1	2	3	2	1	3	2	3
VITALS	0	0	2	2	2	2	2	2	2	2	2	2	1	1	1	1	2	1	2	1	0	0
DOCTORS	0	0	4	4	4	4	4	4	3	4	2	4	2	1	2	2	1	4	0	1	1	0
BLOOD TESTS	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
X-RAYS	0	0	2	2	2	2	2	2	2	2	2	2	2	2	1	2	2	1	0	1	0	0
PHARMACY	2	2	2	2	2	2	2	2	2	2	1	2	2	1	2	1	2	1	2	1	1	0

Objective function 3

Process	7h30	8h	8h30	9h	9h30	10h	10h30	11h	11h30	12h	12h30	13h	13h30	14h	14h30	15h	15h30	16h	16h30	17h	17h30	18h
CLERKS	3	2	3	3	1	3	2	3	3	1	2	3	2	2	1	3	2	3	2	3	3	3
VITALS	0	0	2	2	2	2	2	2	2	2	2	2	1	2	1	1	1	2	1	1	0	0
DOCTORS	0	0	4	4	4	4	4	3	4	3	4	2	2	1	2	1	1	2	1	2	0	3
BLOOD TESTS	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
X-RAYS	0	1	2	2	2	2	2	2	2	2	2	2	2	1	1	1	2	1	1	1	0	0
PHARMACY	2	1	2	2	2	2	2	2	1	2	2	2	2	1	2	1	2	2	1	1	2	0

TABLE 7.2: *The OPD staff schedules generated by the genetic algorithm for the 2015 OPD set-up.*

Stellenbosch University https://scholar.sun.ac.za

Stellenbosch University https://scholar.sun.ac.za

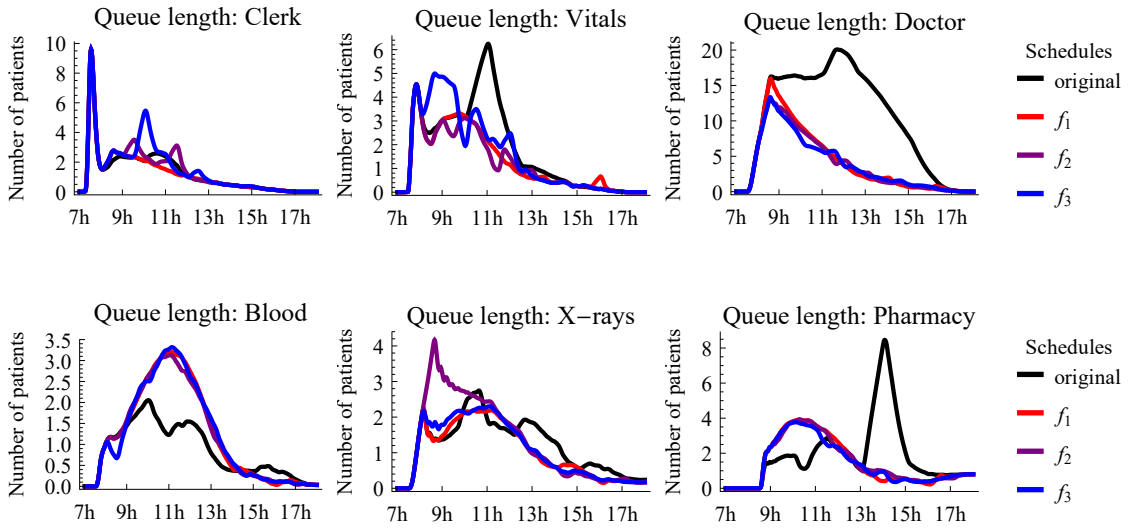


FIGURE 7.5: *A comparison of the average queue length at each OPD process using the OPD staff schedules generated by the genetic algorithm.*

schedules all result in very similar queueing patterns over the course of the day. The most significant difference between the simulation results for the new schedules and the original schedule is the decrease in the length of the DOCTORS queue. By assigning a higher number of doctors during the morning, the new schedules are able to avoid the backlog of patients that occurs between 9h00 and 13h00 in the original set-up.

The new schedules have very little impact on the CLERKS queue, although there are occasional spikes in the average queue length during periods where there are fewer staff on duty in the second and third schedules. This also occurs in the VITALS queue, although the new schedules do remove the mid-morning peak which was caused by the tea breaks in the original schedules. The BLOOD TESTS, X-RAYS, and PHARMACY queues are slightly longer during the morning and early afternoon due to the increased efficiency of the DOCTORS queue, and the PHARMACY queue is slightly longer in the morning and much shorter during the afternoon.

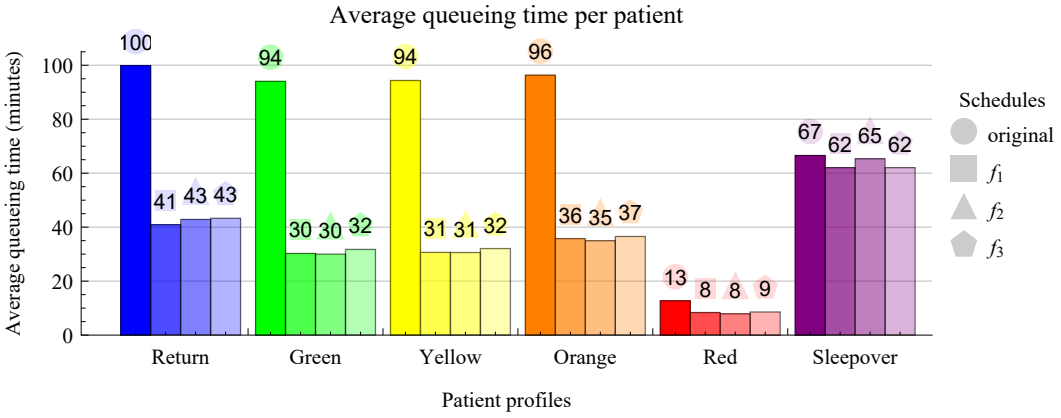


FIGURE 7.6: *A comparison of the average total waiting times for different patient profiles using the OPD staff schedules generated by the genetic algorithm.*

Stellenbosch University https://scholar.sun.ac.za

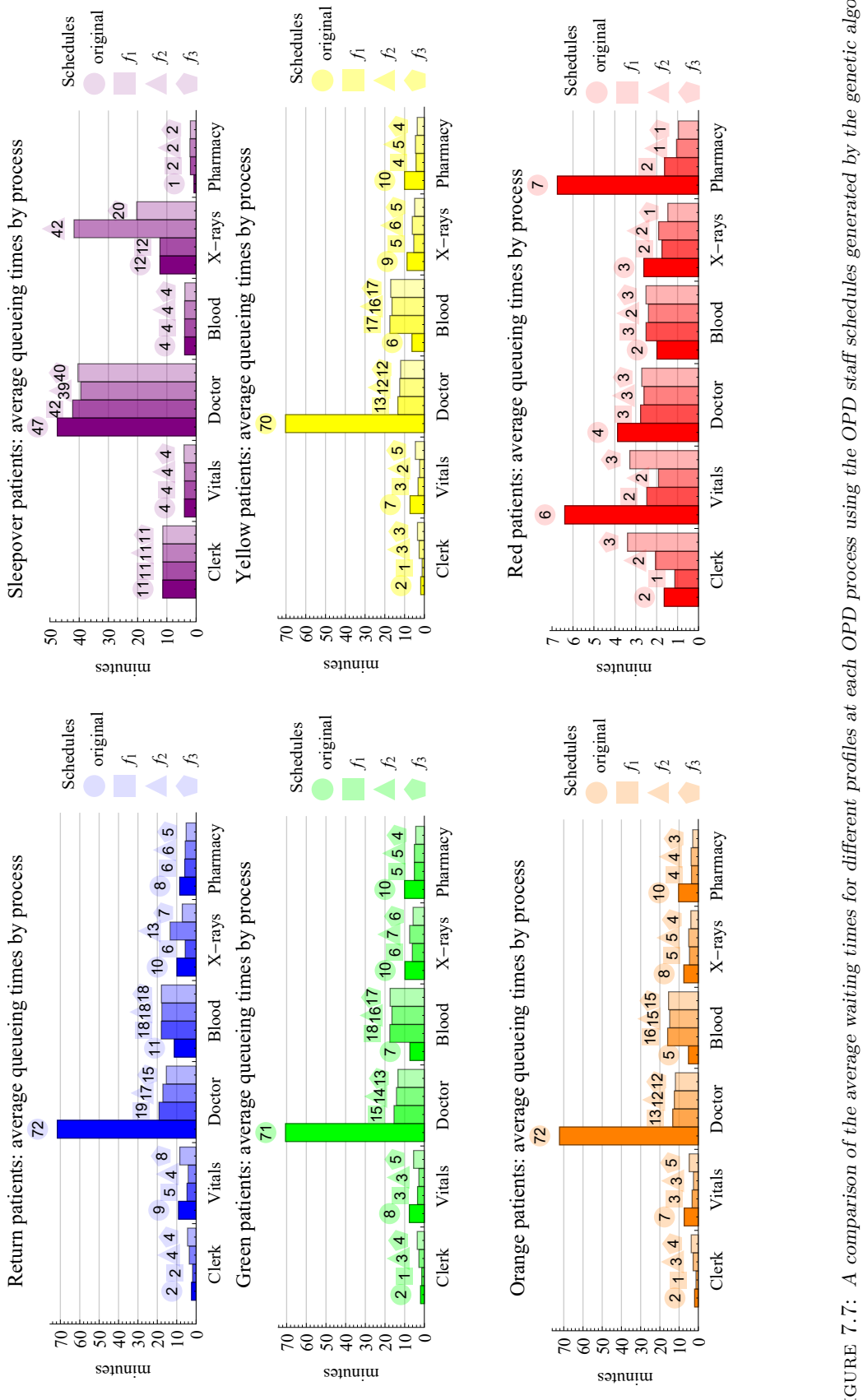


FIGURE 7.7: *A comparison of the average waiting times for different profiles at each OPD process using the OPD staff schedules generated by the genetic algorithm.*

Stellenbosch University https://scholar.sun.ac.za

Figure 7.6 shows that the new schedules produce significantly better average waiting times than the original schedule. The waiting times for return, green, yellow and orange patients all decrease by 60–70%, and the average waiting times for red patients are 30–40% lower. There is not very much difference in the waiting times for sleepover patients, due to the fact that doctors only begin working at 8h30 in all four of the schedules.

Figure 7.7 illustrates the average delays for the different patient profiles at each OPD process. The new schedules decrease the average amount of time that return, green, yellow and orange patients spend in the DOCTORS queue by 75–85%, and there is also a small reduction (5–8 minutes) in the amount of time that sleepover patients spend in this queue. The waiting times at the BLOOD TESTS queue increase by 8–10 minutes for the first four profiles, which is due to the increased number of casualty patients leaving the DOCTORS queue in the mornings.

The three new schedules result in very similar average waiting times at each process, but there are a few instances where their performance differs. Schedule 3 performs slightly worse than the other two schedules in terms of the average waiting times at the CLERKS and VITALS queues, but there is still an improvement in the waiting times at these processes relative to the original schedule. Schedules 2 and 3 have significantly higher waiting times for sleepover patients in the X-RAYS queue — 42 minutes and 20 minutes, compared to 12 minutes for the original schedule. This does not occur with the first schedule, which has a higher number of staff at this process at the beginning of the day.

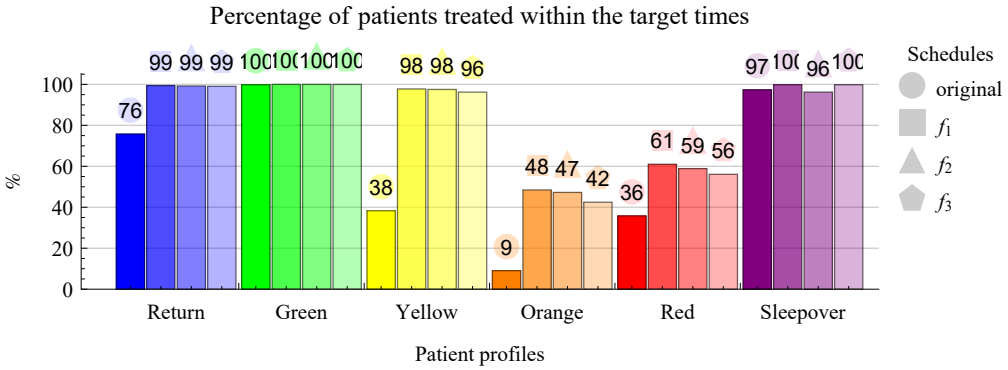


FIGURE 7.8: *A comparison of the number of patients treated within the target times using the OPD staff schedules generated by the genetic algorithm.*

The chart in Figure 7.8 illustrates the percentage of patients treated within the appropriate target times for each of the patient profiles. The new schedules all result in similar numbers of missed targets and perform better than the original schedule, particularly for casualty patients. The number of patients treated within the target time increases for yellow patients (38% to 96–98%), orange patients (9% to 42–48%), red patients (36% to 56–61%) and return patients (76% to 99%).

7.7 Conclusion

Based on the results presented in § 7.6, adjustments to the OPD staff schedules can help to reduce the level of congestion in the facility during busy periods. The most efficient staff schedules are aligned with the patient arrival rates, which tend to be highest in the mornings and lower in the afternoons.

Stellenbosch University https://scholar.sun.ac.za

The number of doctors on duty at different times during day has the biggest impact on the flow of patients through the OPD, since most patients spend more time in the DOCTORS queue than in any of the other queues. Scheduling additional doctors during the morning helps to avoid a long backlog of patients during the middle of the day and leads to shorter waiting times.

Although the schedules presented in § 7.6 could be feasibly implemented, they would require some staff to work short, fragmented shifts with irregular hours, or to go long periods without a break. These schedules would also be problematic for the OPD doctors, who often work at local clinics or in other parts of the hospital during the mornings. Additional constraints are needed in the genetic algorithm to account for these restrictions and to avoid producing schedules that involve too many staff changes.

Stellenbosch University https://scholar.sun.ac.za

Stellenbosch University https://scholar.sun.ac.za

CHAPTER 8

Decision support tool

This chapter describes a decision support tool that was built for the OPD. The purpose of this software is to provide insight into the queueing process in the OPD based on the results of the OPD simulation model. Using this software, the hospital staff can gain a better understanding of the factors that contribute to delays in the OPD and assess different strategies for improving the efficiency of the OPD.

The first part of this chapter (§ 8.1) considers the motivation for the decision support tool and explains why this approach is the most useful, practical way to implement the OPD model. An overview of the decision support tool software is provided in § 8.2, and § 8.3 discusses feedback regarding the implementation of the decision support tool.

8.1 Motivation

One of the aims of this project is to make a practical contribution to the ongoing efforts to improve the Zithulele OPD. To achieve this, the implementation of the OPD simulation model must compliment these efforts and be easily integrated with the long term initiatives that are already under way. The OPD decision support tool seeks to bridge the gap between this (academic) project and the practical initiatives at Zithulele.

The OPD app is based on the simulation model in Chapter 4. The application provides a framework for staff to create their own simulation set-ups, adjust the input parameters, run the simulation model, and view the results. Based on these results, staff members can assess the efficiency of their OPD set-ups and compare the results of different systems.

This approach to evaluating and improving healthcare systems was introduced by Fetter & Thompson (1965), who give a very clear description of the purpose of such models:

> "Our objective is to provide hospital administrators with tools that will give them the ability to predict the operational consequences of designs and, given any set of facilities, the results of the application of alternative policies for guiding the operation of these facilities."

An important aspect of this approach is the emphasis on empowering hospital staff to make more informed decisions, rather than building models which attempt to make these decisions for them. This is particularly relevant in this project, because changes to the Zithulele OPD

Stellenbosch University https://scholar.sun.ac.za

are more likely to be implemented if they come from within the hospital, rather than external recommendations.

Throughout this project, the OPD app played a key role in communications with the staff at Zithulele and the development of the OPD simulation model. Using the app, the OPD staff were able to test different versions of the model and provide feedback about how it could be improved. Both the app and the underlying simulation model were expanded several times to incorporate these ideas and suggestions.

The earlier versions of the OPD app were very simple and focussed on the accuracy of the underlying model. Later in the project, the focus shifted to building a useful and informative decision support tool that could be used by the staff at Zithulele. This process was based on four important criteria:

1. The decision support tool must provide ongoing feedback.

   As in any other large organisation, changes to the Zithulele OPD system cannot occur instantly. The improvements to the OPD are a continuous process of adjustments and re-evaluation, so a once-off analysis of the system is not very useful. The decision support tool therefore needs to be available to the hospital on a long term basis.

2. The model must be flexible.

   The decision support tool must be adaptable to changes that might occur at Zithulele over the next few years. This might include new treatments, changes in the number/composition of patients and the availability of staff and equipment. Ideally, the decision support tool should also be flexible enough to be used at similar facilities in other hospitals.

3. The decision support tool must be easy to use.

   Since the decision support tool is intended to be used by the hospital staff, it should not require any advanced knowledge of simulation or statistics. It must be detailed enough to produce useful and informative results, but not complicated or confusing from a user's perspective.

4. The results must be presented in a clear, user-friendly format.

   The results provided in the decision support tool should give the user all the information that they need in a simple, logical manner. The results should include summaries which enable the user to gauge the overall efficiency of a system very quickly, as well as detailed information to allow in-depth analysis of the simulation results. All of these results must be presented in a simple, clear format that aids the user's understanding and avoids the potential for misinterpretation.

8.2 The OPD app

The OPD decision support tool was implemented in Python, using PyQt for the user interface. Although this software is open-source and freely available online, the decision support tool is packaged as a standalone app so that the hospital staff can use it without the inconvenience of downloading and installing additional software.

The main window of the app consists of three separate tabs, and users can navigate between the tabs in the same way that they would move between multiple tabs in a browser. The first tab is used to build models, while the second and third are used to view results. These tabs are discussed in greater detail in § 8.2.1–8.2.3.

Stellenbosch University https://scholar.sun.ac.za

The optimisation algorithm from Chapter 7 is also included in the OPD app, but is contained in a separate window which can be accessed from the data tab in the main app. The optimisation interface is discussed in § 8.2.4.

8.2.1 Data tab

The data interface contains all the information needed to run a simulation and change the simulation parameters. Users can save, open and edit OPD set-ups, which provides a convenient way to compare the results of multiple different set-ups. Each saved set-up consists of a single file, so set-ups can easily be shared with other users.

A screen shot of the data tab is shown in Figure 8.1. The tab is divided into three different sections, which are intended to guide users through the process of building a new set-up. The first step is to enter the process data (the top section), followed by the patient profiles (bottom left) and the treatment data (bottom center).

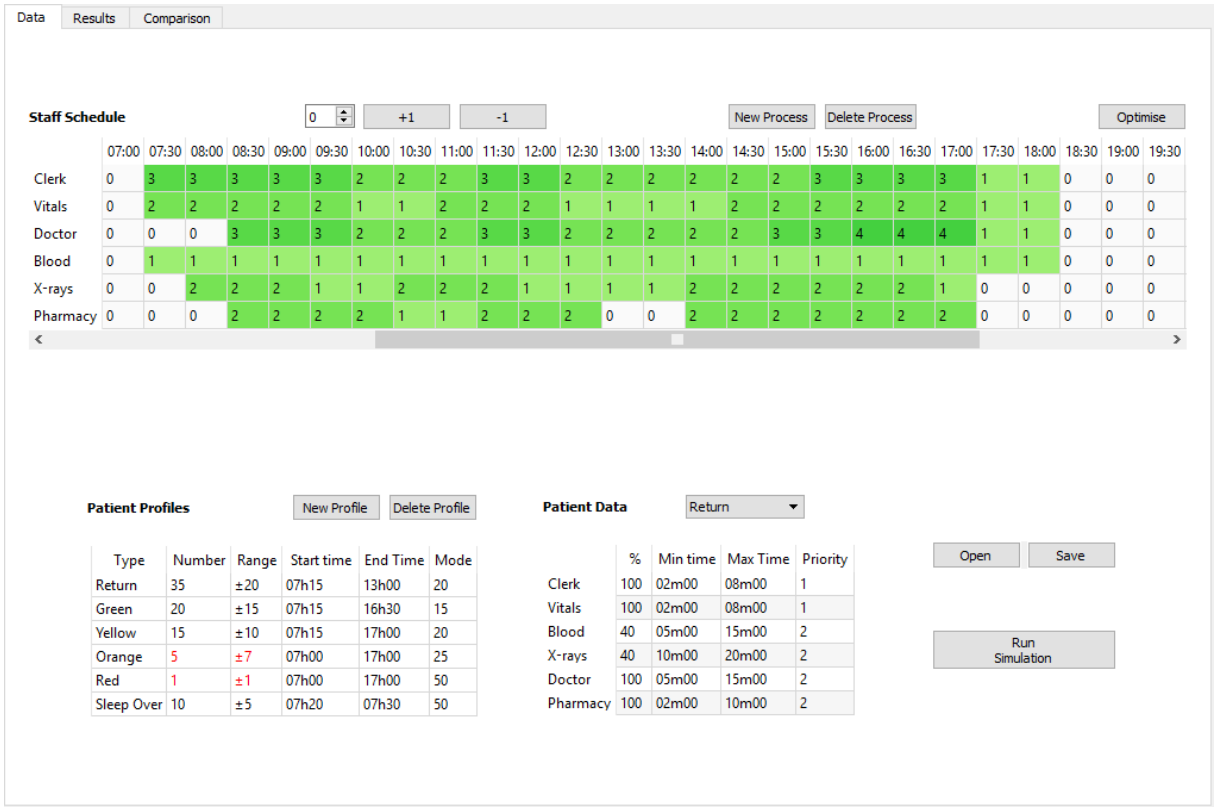


FIGURE 8.1: *A screen shot of the data tab in the OPD app.*

Process data

Processes can be added, removed and renamed in the table at the top of the data tab, which also contains the staff schedule for each process. The staff schedule table is divided into 30 minute increments and cells are shaded according to the number of staff on duty to make it easier to read and edit this data.

Each model must have at least one process, but there is no upper limit to the number of processes that can be added. Every process requires at least one staff member on duty at some stage during

Stellenbosch University https://scholar.sun.ac.za

the day and processes that do not meet this requirement are highlighted to indicate that the data is incomplete.

Patient profiles

Users can add, delete and rename patient profiles in the table in the bottom left-hand corner of the data interface. The arrival parameters for each profile in the simulation model are based on the information in this table.

The first two columns of the patient profiles table give the expected number of patients for each profile and a symmetric upper and lower bound for this number. The number of patients in each simulation is generated from a symmetric triangular distribution over this interval.

The last three columns describe the arrival patterns for each patient profile over the course of a day. Random patient arrivals are generated using a triangular distribution between the starting time and the ending time. The mode column allows the user to set the peak of the distribution using a value between 0 and 100, where 0 is the start time, 100 is the end time and 50 is the midpoint of the interval.

Treatment data

The treatment data table at the bottom of the data interface allows users to enter information about how different types of patients interact with the different processes. Users select a particular profile to edit on the drop-down menu above the table, or by clicking on the profile in the profile table.

The treatment data table lists each of the processes in the model and the corresponding treatment parameters for the selected profile. Users can re-arrange the processes to reflect the order in which they are visited by patients from each profile. When these changes are made, they only apply to the selected profile, so different types of patients can complete processes in a different order.

The first column in the table is restricted to values between 0 and 100, which indicate the percentage of patients in the selected profile that need each process. The second and third columns give the minimum and maximum treatment times at each process and the final column gives the priority of the selected profile at each process.

FIGURE 8.2: *A screen shot of the priority window.*

The first three columns in the treatment data table can all be edited by typing in the appropriate value, but the priority parameters are adjusted by dragging the different patient profiles into the

Stellenbosch University https://scholar.sun.ac.za

appropriate order in a separate window (see Figure 8.2). This is more convenient than typing the priorities into the table, because users can change the priority of a particular profile without needing to manually adjust the values for all of the other profiles. For example, if the user drags a specific profile to the top of the list, the corresponding profile is assigned a priority of 1 and the priority of the other profiles is automatically adjusted to reflect their position in the new list.

8.2.2 Results interface

The results tab (Figure 8.3) allows users to view detailed simulation results based on the current set-up in the data tab. Most of the results are displayed in graphs and colour-coded according to the patient profiles, which makes it easy to compare how efficiently each profile was treated.

FIGURE 8.3: *A screen shot of the results tab in the OPD app.*

Statistics

Basic waiting time statistics are provided in the table on the right of the results tab. For each profile, the number of patients is given, followed by the total waiting time, total time spent in the OPD (in hours) and the ratio of waiting time to treatment time. These values are given as an average per patient, followed by the minimum (best) and maximum (worst) values.

Simulation playback

The controls at the top of the results tab allow the user to view an animated playback of the simulated queues in the OPD. Users can play, pause, and adjust the speed of the animation or

Stellenbosch University https://scholar.sun.ac.za

skip through the simulation one event at a time. A description of each event is provided at the top of the screen to indicate which patient has moved, as well as a few basic details about how long they queued for that particular treatment. The time of the event is displayed to the left of this text and the four graphs at the bottom of the results tab illustrate the state of the OPD when the event occurred.

The first graph (labelled "Queue Length") shows the number and profile of patients in the queue at each process. The patients currently waiting in that queue are represented by a series of rectangles which are colour-coded according to the patient profiles. A thin green bar appears over patients at the front of the queue to indicate that they are currently being treated by a staff member. The graph is updated every time a patient joins a different queue or begins a new treatment.

The second graph (labelled "Total Queueing/Treatment Hours") shows the accumulation of waiting time for each of the different processes over the course of the day. Different colours are again used to show how this waiting time is distributed among the different patient profiles. A smaller green bar indicates the accumulation of treatment time at each process.

The graph in the bottom left corner (labelled "Arrival/Exit") illustrates the total amount of time that each patient spent in the OPD by a single line (colour-coded by profile). The bottom of the line indicates when the patient arrived, and the top of the line shows when the patient left.

The last graph (labelled "Treatment/Waiting Hours") provides a more detailed breakdown of the waiting to treatment time ratio. Each line in this graph represents the total amount of time that a patient spent in the OPD. The part of the line below the black circle shows how much of this time was spent being treated by a staff member, while the part of the line above the circle represents the amount of time spent queueing.

The visual nature of the results in these graphs makes it easy to evaluate the overall efficiency of the OPD system in that simulation, to identify long delays, and to investigate why they occurred. By combining the information from the graphs and statistics, users are able to see how different patients and processes interact with the rest of the system.

8.2.3 Comparison interface

The aim of the comparison tab is to allow users to view the results of multiple simulation runs at the same time. This is very important in order to avoid being misled by a single set of very poor or very promising results. It also helps to give users an idea of how stable and predictable the current set-up might be. Although the hospital is most concerned with worst-case scenarios (i.e. days when there are very long queues), it is also important to look for relatively stable configurations where the state of the hospital is less likely to vary dramatically from day to day.

The controls on the left of this interface allow the user to specify how many simulations to run and to choose which processes they would like to view in the results. Existing results are not deleted unless the user clears the graphs, so the results of new simulations are added to the current graphs.

This is very useful if the user would like to look at the effect of changes to the system (for example, a new staff schedule or an increase in the number of patients). The user begins by opening the first set-up and performing a few runs of the simulation model with these parameters. They then open the second set-up in the data tab and return to the comparison tab to add new simulation runs using the updated data. The new set-up's results are plotted alongside the previous ones.

Stellenbosch University https://scholar.sun.ac.za

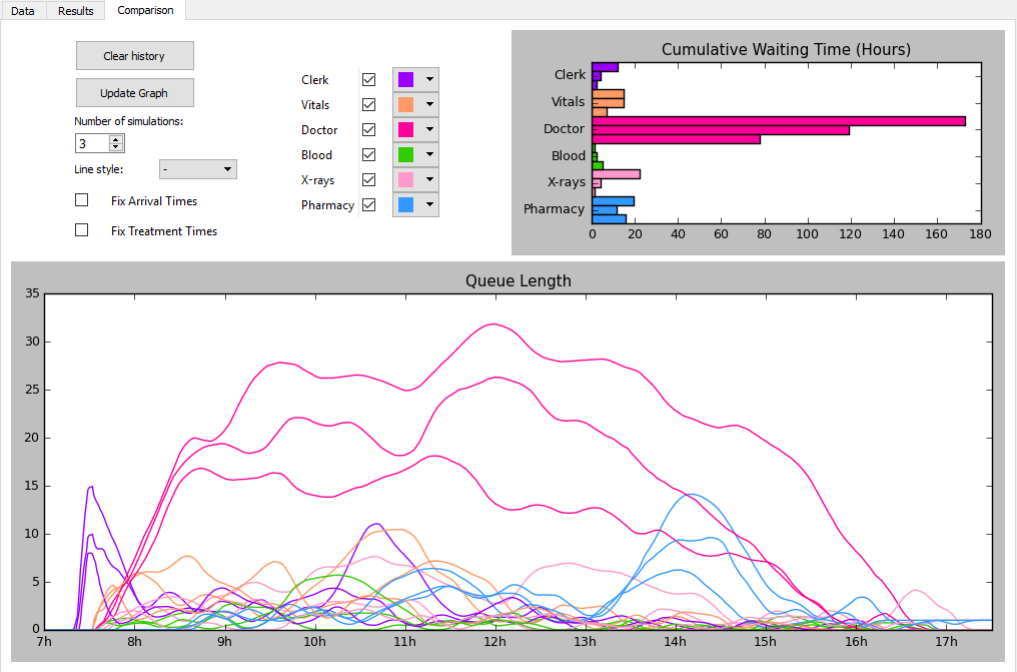


FIGURE 8.4: *A screen shot of the comparison tab in the OPD app.*

The top graph in this window compares the cumulative waiting time at each process across multiple simulations. Users can identify which processes contribute most to the waiting time and how much the waiting time varies from one day to the next.

The bottom graph shows the queue length at each process over the course of the day. Apart from providing an overview of how queue lengths might vary from day to day, this graph is also very useful for illustrating which queues cause the longest delays on a daily basis. In order to make the graph easier to read, the queue length plots are smoothed over five minute intervals.

Viewing multiple queues in the same graph helps to identify interactions between different queues. For example, an influx of patients at a particular process generally has a wave-like effect at processes further down the line, and increasing the staff at a particular process may result in longer queues at other processes.

In the queue length plot, users can select different line styles for new simulations. When the line style is changed, all existing lines stay the same and only new results are plotted using the new style. This is a very useful way to differentiate between the results of two or more set-ups.

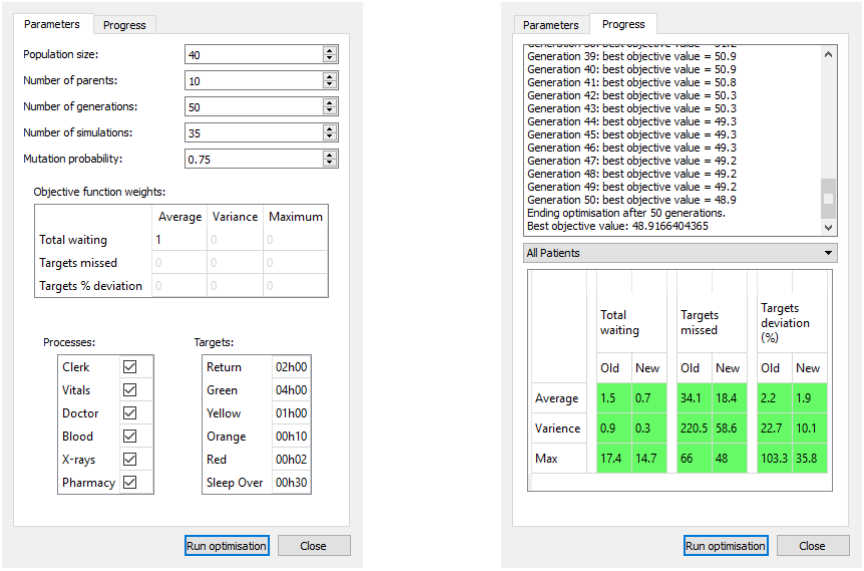
Fixed variables

Above the queue graph, users my choose to fix either the arrival times or the treatment needs of each patient (fixing both will result in identical simulations). This can help to determine which processes are very sensitive to differences in patient arrivals and which are more likely to be affected by variation in the treatment needs of individual patients. It can also identify instances where the data provided is too vague (for example, if the range for treatment times is too wide). Often, this can be addressed by dividing a patient profile into two sub-profiles with a smaller range of treatment times.

Stellenbosch University https://scholar.sun.ac.za

8.2.4 Optimisation interface

The optimisation model from Chapter 7 was not originally part of the decision support tool software due to the difficulties associated with identifying the appropriate optimisation constraints (see § 7.2). However, it was added at the request of the OPD staff, who felt that the optimisation results were interesting even if they were not always practically implementable.



(a) *The optimisation settings interface.*

(b) *Optimisation progress and statistics.*

FIGURE 8.5: *A screen shot of the optimisation interface in the OPD app.*

Parameter tab

The first tab in the optimisation interface (Figure 8.5(a)) allows users to set up the variables, parameters, and fitness function for the genetic algorithm. This information is divided into four sections:

1. Genetic algorithm parameters

   The population size, number of children, number of simulations, and the mutation probability can be adjusted at the top of the parameter tab in the optimisation interface. The user is not expected to understand the effect of these parameters, and the main reason for their inclusion in the interface is to give the user some control over the running time of the algorithm. Detailed guidelines are provided to indicate appropriate ranges for each parameter.

2. Fitness function

   The statistics available for the fitness function are listed in the table below the optimisation parameters. The weights for these statistics are entered into the table and the sum of these weights must add up to 1.

3. Variables

   In the processes table, users can select which schedules they would like to optimise. The remaining schedules are kept constant.

Stellenbosch University https://scholar.sun.ac.za

4. Targets

   The user can enter a target time for each profile on the right of the parameter tab. The target times are used to calculate the objective function statistics.

Progress tab

When the genetic algorithm begins, a progress tab opens in the optimisation interface (Figure 8.5(b)). At the top of the progress tab, the most recent objective value is printed after each iteration of the genetic algorithm. This allows the user to see whether the staff schedules are still improving and to end the algorithm if the iterations are progressing too slowly.

Once the algorithm is complete, the objective function statistics for the original schedule and the best schedule found by the optimisation algorithm are compared in the table at the bottom of the progress tab. These statistics are based on fresh simulations, rather than the simulations used in the genetic algorithm. All nine statistics are shown, regardless of whether they were included in the fitness function, and the cells are shaded green or red to indicate whether they have decreased (improved) or increased.

The old schedule is also compared to the improved schedule in the comparison tab. Three simulations are run using each of the schedules and plotted on the same graphs. The optimised schedule is copied into the data tab so that users can view and save the new set-up.

8.3 Results and feedback

This section discusses some of the insights gained during the design and implementation of the OPD app. The first part of this discussion focusses on how the app was used during the course of this project. Section 8.3.1 explains the value of the OPD app from an academic perspective and describes its role in the mathematical modelling process, while § 8.3.2 describes the practical benefits of the app and the feedback received from the OPD staff. The long term applications of the decision support tool are considered in § 8.3.3.

8.3.1 Academic contributions: modelling the OPD system

The OPD app played an important role in the development of the OPD conceptual model (Chapter 2) and the simulation model (Chapter 4). Using the OPD app, the staff at Zithulele were able to run simulations and provide feedback about how well the simulation model captured the behaviour of the OPD queues, based on their knowledge and experience of the system.

This feedback was very beneficial to the modelling process, especially considering the scarcity of data to describe the OPD queues. Interactions with the OPD staff helped to evaluate the accuracy of the model and the validity of its assumptions. Suggestions from the OPD staff were used to adjust and improve the model on several occasions, which resulted in a greater level of detail in the simulation model.

There were certain challenges associated with using the OPD app in the modelling process, such as the need to limit the simulation model to open-source software. The initial simulation model was coded in Matlab and Mathematica, but had to be re-coded in Python when it was implemented in the app. It was also necessary to resolve compatibility issues associated with different operating systems to ensure that both the app and the underlying model ran smoothly on different computers.

Stellenbosch University https://scholar.sun.ac.za

In addition to these technical considerations, the scope and complexity of the simulation model were somewhat restricted by the need to maintain the simplicity of the OPD app. Since the simulation model was coded in Python, changes to the model were easy to implement. However, as the level of detail in the model increased it became significantly more difficult to incorporate these improvements in the OPD app's user interface.

For example, extending the simulation model to allow priority queue disciplines required $n \times m$ additional input parameters to indicate the priority of each patient profile at each process. It was difficult to include these parameters in the OPD app due to the layout of the data interface, which only displays the treatment parameters for one profile at a time. To allow the user to view and edit the relative priorities for each of the different profiles at the same time, it was necessary to build an additional priority window for the data interface.

If the simulation model had not been subject to the restrictions of the OPD app, it may have been possible to achieve a more detailed and accurate representation of the OPD system. However, this model would not be accessible to the staff at Zithulele. In this regard, the limitations of the OPD app were beneficial because these restrictions helped to maintain an appropriate balance between complexity and accessibility in the underlying model.

8.3.2 Practical contributions

This section describes how the OPD app has contributed to a better understanding of the queueing system at Zithulele. These observations are based on feedback from the OPD clinical manager, Dr Ben Gaunt, as well as meetings and interactions with the OPD staff which took place at Zithulele in April 2015 and December 2015.

Understanding the queueing network

Due to the complexity of the OPD queueing system and the large number of patients in the facility, it is difficult to visualise how these queues function as a network. The OPD app addressed this problem by providing a visual representation of the OPD system that clearly illustrates the flow of patients from process to process. The graphs in the results tab allowed staff to identify specific problems that need to be addressed and to investigate the circumstances that lead to these problems.

The data interface of the OPD app provided insight into how characteristics of the patient profiles and processes influence the efficiency of the OPD queues. Adjusting the parameters in the OPD app allowed staff to develop a better understanding of the different types of parameters, as well as their relevance to the model.

Facilitating communication

The OPD app was a very useful tool to facilitate discussions about the efficiency of the OPD among the OPD staff. This was particularly important, because most staff tend to focus on the queue that they are assigned to treat and are not necessarily aware of the queues at other processes. The graphical results in the OPD app allowed staff to communicate about these different perspectives and to understand how their role in the OPD is linked to the other processes in the network.

Stellenbosch University https://scholar.sun.ac.za

Evaluating strategies

The OPD app allowed users to compare the efficiency of different OPD set-ups, including changes to the processes, staff schedules and treatment parameters. In discussions with the OPD staff, this was used to develop strategies to address specific problems in the OPD, and to determine the most efficient way to distribute resources among the different processes.

When discussing a specific problem in the OPD, proposed solutions often focus on a single process or profile and fail to account for the effect of a particular strategy on other components of the system. For example, assigning additional staff to a busy queue may help to reduce delays at that process, but can result in longer queues at other processes and very little improvement in the total waiting times. It was helpful to test proposed solutions in the OPD app, because the simulation model demonstrates how changes to a certain part of the system can have unexpected consequences in other parts of the network.

The optimisation algorithm in the OPD app was also useful in understanding how to schedule staff more efficiently. Although it does not necessarily produce schedules that are implementable, these improved schedules provide useful guidelines for matching the number of staff on duty to the periods of high and low traffic in the OPD. These results led to a very interesting discussion about the constraints on the doctors' schedules, which are related to their work in other parts of the hospital.

Data awareness

An indirect result of the implementation of the OPD app at Zithulele was an increased awareness of the different types of data that are needed to monitor the OPD system. Many of the treatment parameters in the simulation model were aligned with data collected during the audit, and discussions about these parameters led to a better understanding of the arrival patterns for different patient profiles in the OPD. Cases where certain treatment parameters could not be estimated from the available data were also useful, because these instances highlighted additional information that could be collected in future.

The OPD app also helped to explore different techniques that can be used to measure the efficiency of the OPD. Working with the simulation and optimisation results illustrates the relationship between various efficiency measures such as queue lengths, waiting times and waiting time targets. This helped to formulate more specific descriptions of some of the problems in the OPD by relating them to these outcomes, and to clarify how to measure the impact of changes to the system.

8.3.3 Future applications

Over the next few years, the facilities at Zithulele will be expanded to address the overcrowding in the OPD. The OPD app can be used as a decision support tool for planning these new facilities in order to avoid perpetuating many of the problems that arise in the current set-up. Using the app, staff can identify which parts of the current OPD system need to be improved and investigate how to address these issues. The app can also be used to anticipate future problems that might arise in the new system if the number of OPD and casualty patients continues to increase. Including these considerations in the planning process will help to ensure that the new facilities do not encounter the same issues with overcrowding a few years after they are opened.

The addition of these new facilities is a good opportunity to implement large-scale changes to

Stellenbosch University https://scholar.sun.ac.za

the OPD system. Although it will be challenging to restructure the system over a short period of time, this strategy is likely to yield better results and fewer long term disruptions than a slow series of gradual changes.

The OPD app can also facilitate the implementation of changes by providing a visual representation of how the new OPD system will work. Using the simulation results in the OPD app, hospital management can communicate these ideas in a clear, effective manner and demonstrate how they will be beneficial to both patients and staff. This will help to reduce the resistance to these changes and smooth the transition to the new facilities.

Stellenbosch University https://scholar.sun.ac.za

CHAPTER 9

Conclusion

This chapter provides a brief summary of the work presented in this thesis. Section 9.1 discusses how the aims and objectives of the thesis were achieved, and § 9.2 provides practical recommendations based on the findings of this research. The chapter concludes with a discussion of potential extensions of this research in future work.

9.1 Summary and achievement of objectives

The aims and objectives of this research (introduced in § 1.2) include both academic and practical components. This section outlines how each of these aims and objectives was achieved.

Aim 1: Understanding congestion in the OPD

Objective 1.1: Develop detailed mathematical models to describe patient flow in the OPD.

This objective was achieved through the OPD models outlined in Chapters 2, 3, and 4. The conceptual model in Chapter 2 captures the structure of the OPD system and provides a detailed description of the important components within the system. The conceptual model was implemented using queueing theory and simulation techniques in Chapters 3 and 4.

The fluid models in Chapter 3 focus on the behaviour of the OPD queues. Since these models are continuous approximations of a discrete system, they do not give an accurate description of the lengths of the OPD queues when there are periods of low traffic intensity. However, they do provide a reasonably accurate description of patient flow in the discrete system (see 6.2.1).

The agent-based simulation model in Chapter 4 takes a more patient-centered approach, tracking the experiences of individuals within the OPD system. This detailed information is an effective way to understand both the overall level of congestion in the system as well as its implications for individual patients. Another important feature of the simulation model is its ability to represent the day-to-day variability in patient flow and the length of the OPD queues.

Objective 1.2: Analyse the causes and effects of congestion in the OPD.

Chapter 6 contains an extensive analysis of the congestion in the OPD system, based on the results of the mathematical models in Chapters 3 and 4. These results illustrate important

Stellenbosch University https://scholar.sun.ac.za

factors that affect congestion in the facility, such as interactions between different queues within the OPD network (see § 6.1.5), and the varying effects of congestion on different types of patients (see § 6.1.3).

The results in Chapter 6 indicate that the causes of congestion are linked to the distribution of patient arrival times and the OPD staff schedules (see § 6.3.1). Backlogs during peak arrival periods are exacerbated by low staff levels, especially in the DOCTORS queue. The effects of congestion are most severe for urgent casualty patients, who are likely to experience unacceptably long delays during busy periods (see § 6.3.2).

Aim 2: Strategies for improving the OPD

Objective 2.1: Evaluate strategies to address the causes of congestion.

Two potential strategies to address the causes of congestion are considered in § 6.3.1. The first strategy is to reduce the number of patient arrivals during peak periods by shifting less urgent arrivals to later in the day. Unfortunately, this strategy is unlikely to be practically implementable due to the location of the hospital.

The second strategy is to align the OPD staff schedules with the variations in patient arrivals over the course of the day. This strategy is investigated through an optimisation model in Chapter 7. The optimisation results in § 7.6 show this strategy is a very effective way to decrease congestion and delays in the OPD.

Objective 2.2: Evaluate strategies to mitigate the negative effects of congestion.

Strategies to address the negative effects of congestion are discussed in § 5.3.2, which describes changes that were implemented in the OPD during the course of this research. These strategies focus on reducing delays for urgent casualty patients by identifying and prioritising these patients in the OPD queues. The analysis of the simulation results in Chapter 6 suggests that these strategies are an effective way to reduce waiting times for patients with urgent medical needs. Section 6.3.2 discusses how prioritisation systems can also help to alleviate overcrowding in parts of the facility and reduce the number of patients who wait overnight at the OPD.

Aim 3: Practical implementation

Objective 3.1: Develop a decision support tool to give hospital staff access to models and results.

The OPD simulation model has been successfully implemented as a decision support tool, which is described in Chapter 8. The decision support tool provides useful insights into the queueing process and allows staff to assess the effect of potential changes to the OPD system. It has already been used during the course of this project (see § 8.3), and its high level of flexibility will ensure that it remains useful in the long term.

9.2 Recommendations and conclusions

This research has emphasised the importance of understanding the complex network of interactions that take place in the OPD. Awareness of congestion in the OPD is generally linked to

Stellenbosch University https://scholar.sun.ac.za

visible signs, such as the length of the queues at specific OPD processes. Although these visible signs may indicate immediate problems that need to be addressed, they should be considered in the context of the entire system. Efforts to improve the efficiency of the OPD system need to address the overall flow of patients through the network, rather than targeting specific problems in isolation.

Concerns regarding the effect of congestion on specific types of patients are also very important, since the dynamics of the OPD queues affect different patient profiles in different ways. It is particularly challenging to understand these dynamics when many profiles are mixed together in long queues, because the composition of these queues can shift over the course of the day. An increased awareness of the arrival patterns for different types of patients can help to clarify how these patients are affected by busy periods.

Based on the results presented in this thesis, there are a number of ways to reduce the negative effects of congestion in the OPD within the scope of Zithulele's limited resources. Scheduling strategies, such as redistributing some of the afternoon staff to morning shifts, can help to avoid long backlogs and reduce the amount of time that patients spend in the OPD.

Priority queues are an effective way to balance the needs of different patient profiles within the OPD system. This strategy can be used to avoid long delays for urgent patients and to limit the variability in these waiting times during periods of high and low congestion. Prioritising the queues that tend to be very busy during the morning also allows patients who need to visit several other processes to move through the system more quickly, and reduces the number of patients waiting at the OPD overnight.

The hospital has already taken steps towards understanding and improving the efficiency of the OPD through the 2015 audit and the implementation of separate queues for casualty and returning patients. Long term improvements to the efficiency of the OPD will require ongoing monitoring and assessment of the system, and the hospital should therefore consider strategies for including continuous data collection in their general patient records and administrative procedures.

9.3 Future work

This section explores opportunities for future work related to the problems considered in this thesis. Refinements and improvements to the existing models are discussed in § 9.3.1, and potential extensions of the scope and applications of this research are suggested in § 9.3.2.

9.3.1 Depth extensions

The most challenging aspect of this thesis was the lack of reliable data concerning patients and staff in the OPD. Reasonable steps were taken to validate the OPD models through discussions with staff members, but the accuracy of these models could be improved in a number of ways if additional data becomes available.

1. Improved parameter estimates

   The sensitivity analysis in § 6.1.5 highlights the importance of the treatment parameters in the OPD simulation model, and indicates that the hospital should focus on collecting information regarding the DOCTORS queue. These parameters can be updated in the decision support tool to improve the accuracy of its results.

Stellenbosch University https://scholar.sun.ac.za

2. Model assumptions

   Additional data regarding treatment times could also be used to investigate the assumption that staff work at a constant speed, regardless of the level of congestion in the facility. If this assumption proves false, the existing OPD models could be adapted to include these variations in treatment times.

3. Weekday variations

   An issue that was not considered in this research is the variation in patients arrivals on different days of the week. Data collected during the audit indicates that higher numbers of patients tend to come to the OPD on Monday than on any other day of the week, and that Tuesday and Thursday are generally busier than Wednesday or Friday. Including these trends in the OPD model would make it easier to anticipate congestion in the facility and investigate whether different efficiency strategies are appropriate for specific days.

4. Optimisation

   Expanding the OPD model to incorporate weekday variations could also be useful in the optimisation model. If there are significant discrepancies between patient arrivals on certain days, a uniform daily staff schedule will not perform well. Weekly staff schedules would provide a greater scope for optimisation, and also allow for the inclusion of more detailed constraints regarding staff availability on specific days.

9.3.2 Breadth extensions

Although this research focusses on the Zithulele OPD, the decision support tool developed for this facility could potentially be used in other hospitals that have a similar system in their casualty/out patient departments. Additional work would be required to **(a)** investigate whether any adjustments should be made to the underlying simulation model; **(b)** ensure that the user interface is compatible with the hardware/software at other hospitals; **(c)** train staff to use the decision support tool; and **(d)** assist with data collection and analysis to determine parameter estimates.

Another potential avenue for future work is the development of a broader model which includes the in-patient wards and other hospital facilities. This would provide a more comprehensive understanding of the factors that influence the efficiency of the OPD, since certain staff and resources are shared by different parts of the hospital and patients are often transferred between the different departments within the hospital.

Finally, it may be useful to develop statistical models that describe how waiting times and congestion influence the level of care provided to patients in the OPD. The analysis presented in this thesis assumes that longer waiting times are associated with poorer care and outcomes, particularly when the target waiting times are exceeded. More detailed analysis could be achieved through further research into the relationship between increased waiting times and patient outcomes.

Stellenbosch University https://scholar.sun.ac.za

References

Agnew CE, 1976, *Dynamic modeling and control of congestion-prone systems*, Operations Research, **24(3)**, pp. 400–419.

Albin SL, 1982, *On Poisson approximations for superposition arrival processes in queues*, Management Science, **28(2)**, pp. 126–137.

Baleta A, 2009, *Rural hospital beats the odds in South Africa*, The Lancet, **374(9692)**, pp. 771–772.

Banks J, Carson J, Nelson B & Nicol D, 2004, *Discrete-Event System Simulation*, 4th Edition, Prentice Hall.

Bateman C, 2013a, *Drug stock-outs: Inept supply-chain management and corruption*, South African Medical Journal, **103(9)**, pp. 600–602.

Bateman C, 2013b, *Leadership, commitment and core values garner national award*, South African Medical Journal, **103(10)**, pp. 707–708.

Bertscher A, *Improving patient flow and reducing waiting times at Zithulele Hospital Out Patients Department*, research report, University of Cape Town.

Bertsimas D & Mourtzinou G, 1997, *Multiclass queueing systems in heavy traffic: An asymptotic approach based on distributional and conservation laws*, Operations Research, **45(3)**, pp. 470–487.

Bitran GR & Tirupati D, 1988, *Multiproduct queueing networks with deterministic routing: Decomposition approach and the notion of interference*, Management Science, **34(1)**, pp. 75–100.

Borshchev A & Filippov A, 2004, *From system dynamics and discrete event to practical agent based modeling: reasons, techniques, tools*, Proceedings of the 22nd International Conference of the System Dynamics Society, volume 22.

Breier M, 2007, *The shortage of medical doctors in South Africa: Scarce and critical skills research project*, in *Human Sciences Research Council (HSRC) study: A multiple source identification and verification of scarce and critical skills in the South African labour market*. Commisioned by Department of Labour, South Africa.

Cabrera E, Taboada M, Iglesias ML, Epelde F & Luque E, 2011, *Optimization of healthcare emergency departments by agent-based simulation*, Procedia Computer Science, **4**, pp. 1880–1889.

Chan WC, 2014, *An elementary introduction to queueing systems*, World Scientific.

Stellenbosch University https://scholar.sun.ac.za

Chen H & Yao DD, 2013, *Fundamentals of queueing networks: Performance, asymptotics, and optimization*, volume 46 of *Stochastic Modelling and Applied Probability*, Springer Science & Business Media.

Cobham A, 1954, *Priority assignment in waiting line problems*, Journal of the Operations Research Society of America, **2(1)**, pp. 70–76.

Coovadia H, Jewkes R, Barron P, Sanders D & McIntyre D, 2009, *The health and health system of South Africa: historical roots of current public health challenges*, The Lancet, **374(9692)**, pp. 817–834.

Day C & Gray A, 2016, *Health and related indicators*, in Padarath A, King J, Mackie EL & Casciola J (Eds), *South African health review 2016*. Health Systems Trust.

Fetter RB & Thompson JD, 1965, *The simulation of hospital systems*, Operations Research, **13(5)**, pp. 689–711.

Gaunt CB, 2010, *Are we winning? Improving perinatal outcomes at a deeply rural district hospital in South Africa*, South African Medical Journal, **100(2)**, pp. 101–104.

Green L & Kolesar P, 1991, *The pointwise stationary approximation for queues with non-stationary arrivals*, Management Science, **37(1)**, pp. 84–97.

Green L, Kolesar P & Svoronos A, 1991, *Some effects of nonstationarity on multiserver Markovian queueing systems*, Operations Research, **39(3)**, pp. 502–511.

Green LV, Kolesar PJ & Soares J, 2003, *An improved heuristic for staffing telephone call centers with limited operating hours*, Production and Operations Management, **12(1)**, pp. 46–61.

Green LV, Kolesar PJ & Whitt W, 2007, *Coping with time-varying demand when setting staffing requirements for a service system*, Production and Operations Management, **16(1)**, pp. 13–39.

Health Systems Trust, *Health statistics*, Available from `http://indicators.hst.org.za/healthstats/134/data`, retrieved 2016-06-19.

Huang J, Carmeli B & Mandelbaum A, 2015, *Control of patient flow in emergency departments, or multiclass queues with deadlines and feedback*, Operations Research, **63(4)**, pp. 892–908.

Ingolfsson A, Akhmetshina E, Budge S, Li Y & Wu X, 2007, *A survey and experimental comparison of service-level-approximation methods for nonstationary $M(t)/M/s(t)$ queueing systems with exhaustive discipline*, INFORMS Journal on Computing, **19(2)**, pp. 201–214.

Jackson JR, 1957, *Networks of waiting lines*, Operations Research, **5(4)**, pp. 518–521.

Jackson JR, 1963, *Jobshop-like queueing systems*, Management Science, **10(1)**, pp. 131–142.

Jaiswal NK, 1968, *Priority queues*, Academic Press, New York.

Kelly FP, 1975, *Networks of queues with customers of different types*, Journal of Applied Probability, **12(3)**, pp. 542–554.

Kelly FP, 1976, *Networks of queues*, Advances in Applied Probability, **8(2)**, pp. 416–432.

Stellenbosch University https://scholar.sun.ac.za

Kivestu PA, 1976, *Alternative methods of investigating the time dependent $M/G/k$ queue*, Doctoral Dissertation, Massachusetts Institute of Technology.

Kleinrock L, 1975, *Queueing systems, Volume I: Theory*, Wiley Interscience.

Kurtz TG, 1970, *Solutions of ordinary differential equations as limits of pure jump Markov processes*, Journal of Applied Probability, **7(1)**, pp. 49–58.

Kurtz TG, 1971, *Limit theorems for sequences of jump Markov processes approximating ordinary differential processes*, Journal of Applied Probability, **8(2)**, pp. 344–356.

Laskowski M & Mukhi S, 2008, *Agent-based simulation of emergency departments with patient diversion*, Proceedings of the Weerasinghe D (Ed), International Conference on Electronic Healthcare, Springer, London, pp. 25–37.

Liu Y & Gong W, 2003, *On fluid queueing systems with strict priority*, IEEE Transactions on Automatic Control, **48(12)**, pp. 2079–2088.

Liu Y & Whitt W, 2011, *A network of time-varying many-server fluid queues with customer abandonment*, Operations Research, **59(4)**, pp. 835–846.

Mayosi BM & Benatar SR, 2014, *Health and health care in South Africa – 20 years after Mandela*, New England Journal of Medicine, **371(14)**, pp. 1344–1353.

Mejia-Quintero C & Escudero-Marin P, 2015, *ABMS & DES for modelling an emergency department*, Available from `http://www1.eafit.edu.co/asr/courses/research-practises-me/2015-2/students/final-reports/CamilaMejiaQuinteroFinalReport.pdf`, retrieved 2016-06-19.

Morris R, 1981, *Priority queuing networks*, Bell System Technical Journal, **60(8)**, pp. 1745–1769.

Newell C, 1982, *Applications of queueing theory*, Monographs on statistics and applied probability, 2nd Edition, Chapman and Hall.

Odoni AR & Roth E, 1983, *An empirical investigation of the transient behavior of stationary queueing systems*, Operations Research, **31(3)**, pp. 432–455.

Reibman A & Trivedi K, 1988, *Numerical transient analysis of Markov models*, Computers & Operations Research, **15(1)**, pp. 19–36.

Robinson S, 1999, *Simulation verification, validation and confidence: a tutorial*, Transactions of the Society for Computer Simulation, **16(2)**, pp. 63–69.

Robinson S, 2004, *Simulation: The Practice of Model Development and Use*, 1st Edition, Wiley.

Sharma S & Tipper D, 1993, *Approximate models for the study of nonstationary queues and their applications to communication networks*, Proceedings of the IEEE International Conference on Communications 1993, volume 1, Geneva, pp. 352–358.

Siebers PO, Macal CM, Garnett J, Buxton D & Pidd M, 2010, *Discrete-event simulation is dead, long live agent-based simulation!*, Journal of Simulation, **4(3)**, pp. 204–210.

Simelela N, Venter WF, Pillay Y & Barron P, 2015, *A political and social history of HIV in South Africa*, Current HIV/AIDS Reports, **12(2)**, pp. 256–261.

Stellenbosch University https://scholar.sun.ac.za

Spearman C, 1904, *The proof and measurement of association between two things*, The American journal of psychology, **15(1)**, pp. 72–101.

Stainsby H, Taboada M & Luque E, 2009, *Towards an agent-based simulation of hospital emergency departments*, Proceedings of the IEEE International Conference on Services Computing 2009, Bangalore, pp. 536–539.

Stolletz R, 2008, *Approximation of the non-stationary $M(t)/M(t)/c(t)$-queue using stationary queueing models: The stationary backlog-carryover approach*, European Journal of Operational Research, **190(2)**, pp. 478–493.

Taboada M, Cabrera E, Iglesias ML, Epelde F & Luque E, 2011, *An agent-based decision support system for hospitals emergency departments*, Procedia Computer Science, **4**, pp. 1870–1879.

Thom A, Dullaart T, Treatment Action Campaign, Section27 (South Africa) & Eastern Cape Health Crisis Action Coalition, 2013, *Death and Dying in the Eastern Cape: An Investigation Into the Collapse of a Health System*, Treatment Action Campaign and Section27.

Thompson GM, 1993, *Accounting for the multi-period impact of service when determining employee requirements for labor scheduling*, Journal of Operations Management, **11(3)**, pp. 269–287.

Tipper D & Sundareshan MK, 1990, *Numerical methods for modeling computer networks under nonstationary conditions*, IEEE Journal on Selected Areas in Communications, **8(9)**, pp. 1682–1695.

White JA, 2012, *Analysis of queueing systems*, Elsevier.

Whitt W, 1983a, *Performance of the queueing network analyzer*, Bell System Technical Journal, **62(9)**, pp. 2817–2843.

Whitt W, 1983b, *The queueing network analyzer*, Bell System Technical Journal, **62(9)**, pp. 2779–2815.

Whitt W, 1991, *The pointwise stationary approximation for $M_t/M_t/s$ queues is asymptotically correct as the rates increase*, Management Science, **37(3)**, pp. 307–314.

Wolfram Research, Inc, *Mathematica*, Version 10.1.

Wolfram Research, Inc, *NDSolve, Wolfram Language & System – Documentation Center*, Available from `http://reference.wolfram.com/language/ref/NDSolve.html?q=NDSolve`, retrieved 2016-04-01.

World Health Organisation, *South Africa: Country health profile*, Available from `http://www.afro.who.int/en/south-africa/country-health-profile.html`, retrieved 2016-06-19.

Young C & Gaunt B, 2014, *Providing high-quality HIV care in a deeply rural setting – the Zithulele experience*, Southern African Journal of HIV Medicine, **15(1)**, pp. 28–29.

Zithulele Hospital website, *Zithulele hospital website*, Available from `http://www.zithulele.org/`, retrieved 2016-06-19.